

In praise of soft science

'Hard' scientists should stop looking down their noses at social scientists, and instead share methods that could help them address pressing societal problems.

It is the conventional wisdom in the biological and physical sciences, and within research agencies, that the social sciences are, well, 'soft', and lacking in methodological rigour. That's why it is arresting that the US National Science Foundation's prestigious Alan T. Waterman award for young scientists has gone this year to Dalton Conley, a sociologist at New York University (see page 1024).

Conley specializes in the detailed study of economic outcomes among underprivileged groups, and says he avoids research on attitudes because they can't be measured accurately. Research agencies in the United States and elsewhere need to support more social scientists like him, because their work can potentially make a valuable contribution to the study of important societal problems.

Take, for example, climate change and biodiversity loss, two global environmental problems for which human behaviour is a significant driver. Research on these issues tends to focus on the physical nature of the phenomena in question. Study of whatever underlies the behaviour itself is too often regarded as 'soft' science and dismissed as second-rate.

Or consider the relationship between biomedical research and public health. The United States has constructed a well-funded and carefully calibrated system to research and develop the best pharmaceuticals and medical equipment that modern science can provide. But what good is that if patients don't take their drugs correctly, or pharmacists routinely misread a doctor's handwriting? A 1999 US Institute of Medicine study found that medical errors — human errors — kill as many as 98,000 people every year, more than the number who die from traffic accidents, breast cancer or HIV/AIDS. Shouldn't psychology and sociology be better harnessed to address this problem?

Even when social science is confident in its assertions, it often feels that it gets no respect from the outside world. Writing in the current

issue of the American Sociological Association magazine *Contexts*, for example, Harold Wilensky, a political scientist at the University of California, Berkeley, says that social scientists have identified specific, practical solutions for problems such as crime prevention and access to health care. But their advice is largely ignored by US policy-makers, Wilensky argues, adding that governments in northern Europe and Japan have a better track record of implementing social scientists' findings.

It can be argued, of course, that social scientists have brought much of this upon themselves. A lot of their work is politically contentious by its very nature, and the spread of what can be loosely termed 'relativism' has reduced their clout. With so many gifted amateurs working their territory, social scientists have a tougher time asserting the unique nature of their expertise than do astrophysicists, for example. Few of us know much about the dynamics of the cosmos, but we all know plenty about human nature — or at least we think we do.

So the onus falls on the social scientists themselves to hone their methods and ensure that they are ready to stand up to external scrutiny. The National Science Foundation has recognized the need to strengthen methodology in the social sciences. Since the terrorist attacks of 11 September 2001, it has also, to its credit, devoted considerable effort to increasing the resources available to its directorate of social, behavioural and economic sciences.

On the campuses, meanwhile, 'hard' scientists need to get over their disdain for their 'soft' colleagues. The study of society can't just be left to poets and politicians. As the almost boundless complexity of physical and biological systems becomes increasingly apparent, along with the pressing need to better understand patterns of human behaviour, now is as good a time as any for a rapprochement between the two wings of the scientific academy. ■

Not-so-deep impact

Research assessment rests too heavily on the inflated status of the impact factor.

Every year at the end of June, scientific publishers' eyes turn to Philadelphia, where the Institute for Scientific Information (ISI) releases a snippet of data that they crave: the impact factor of each journal. In due course, bureaucrats in research agencies will roll the impact figures into their performance indicators, and those scientists who worry about such things will quietly note which journal's number wins them the most brownie points.

Attempts to quantify the quality of science are always fraught with difficulty, and the journal impact factors are among the few

numbers to persist. The result is an overemphasis of what is really a limited metric.

To obtain the latest impact factors, which were released last week, the ISI number-crunchers added the total number of citations from all the monitored journals during 2004 to items in the journal of interest that were published in 2002 and 2003. They then divided that total by the number of 'citable items' — loosely, papers and review articles — that were published in the journal during those same two years.

The impact factor is taken by some administrators as a measure of the typical citation rate for the journal. But for many journals, it isn't 'typical' at all. *Nature's* latest impact factor is 32.2, an increase on last year and a high number that we're proud of, but it's one that merits a closer look.

For example, we have analysed the citations of individual papers



in *Nature* and found that 89% of last year's figure was generated by just 25% of our papers.

The most cited *Nature* paper from 2002–03 was the mouse genome, published in December 2002. That paper represents the culmination of a great enterprise, but is inevitably an important point of reference rather than an expression of unusually deep mechanistic insight. So far it has received more than 1,000 citations. Within the measurement year of 2004 alone, it received 522 citations. Our next most cited paper from 2002–03 (concerning the functional organization of the yeast proteome) received 351 citations that year. Only 50 out of the roughly 1,800 citable items published in those two years received more than 100 citations in 2004. The great majority of our papers received fewer than 20 citations.

These figures all reflect just how strongly the impact factor is influenced by a small minority of papers — no doubt to a lesser extent in more specialized journals, but significantly nevertheless. However, we are just as satisfied with the value of our papers in the 'long tail' as with that of the more highly cited work.

The citation rate of our papers also varies sharply between disciplines. Many of *Nature's* papers in immunology published in 2003 have since received between 50 and 200 citations. Significant proportions of those in cancer and molecular and cell biology have been in the 50–150 range. But papers in physics, palaeontology and

climatology typically achieved fewer than 50 citations. Clearly, these reflect differences in disciplinary dynamics, not in quality.

The impact factor also mixes citations to diverse types of content: unsurprisingly, review articles are typically the most highly cited, but citations of our Commentaries, News Features and News & Views articles also contribute in a minor way to the numerator (although these items are not counted in the denominator).

The net result of all these variables is a conclusion that impact factors don't tell us as much as some people may think about the respective quality of the science that journals are publishing. Neither do most scientists judge journals using such statistics; they rely instead on their own assessment of what they actually read.

None of this would really matter very much, were it not for the unhealthy reliance on impact factors by administrators and researchers' employers worldwide to assess the scientific quality of nations and institutions, and often even to judge individuals. There is no doubt that impact factors are here to stay. But these figures illustrate why they should be handled with caution. ■

"Impact factors don't tell us as much as some people think about the quality of the science that journals are publishing."

Toyota on a roll

Japan's approach to industrial innovation may be out of fashion, but it still delivers the goods.

For the Japanese car company Toyota, 2005 has been a bumper year. The company's global fortunes are at such a high level that the chairman, Hiroshi Okuda, suggested back in April that it might raise its prices to give "some breathing space" to its bloated US rivals, Ford and General Motors.

The car industry isn't quite the economic driver that it was a few decades ago, but cars still account for a huge portion of consumer spending. And despite the industry's traditional conservatism, it has become a hotbed of innovation in electronics, materials, environmental engineering and other spheres.

Basic scientific research usually operates a few steps away from technological innovation in the motor industry. But Toyota is doing some interesting things at its central research and development laboratory near Nagoya (see page 1026). As in other sectors, the period of transition from scientific knowledge to industrial application is shrinking.

Toyota's success has always been more about industrial efficiency than technical innovation. But its technology has progressed steadily over the past ten years, while the competition in the United States has been resting on its laurels. Now that the boom in sales of large, conservatively designed sport utility vehicles seems to be over, US car-makers are experiencing a rude awakening.

As the prospects for the Detroit industry darkened earlier this year, credit agencies humiliated Ford and General Motors by reducing the ratings of some bonds that the two companies have issued

to 'junk' status. That apparently prompted Okuda's intervention: the Toyota chairman felt that a fresh crisis in the US car industry could lead to a surge in protectionist sentiment that might damage Toyota. Then, a few weeks ago, General Motors announced that it is planning to shed 25,000 people, almost a quarter of its factory workforce in North America.

Perhaps the starkest difference in approach between Toyota and its US rivals has been the way they tackled environmental innovation. Detroit car executives have acted like parodies of themselves, accepting generous subsidies from the federal government under then President Clinton's much-trumpeted Partnership for a New Generation of Vehicles programme and asking for less regulation in return — as though their participation was doing the taxpayer a favour. That initiative came and went, but when oil prices flew through the roof last year it was the Japanese manufacturers Toyota and Honda — not General Motors or Ford — who were ready at the starting gate with their ultra-economic 'hybrid' vehicles.

Toyota's approach to science, technology and innovation isn't exactly off-the-wall. It can't afford to be: the company knows that the product has to work when it is delivered. And Toyota's scientists and engineers don't match the flamboyant modern paradigm of innovation, as inspired by California's Silicon Valley. They are, instead, meticulous, intensely loyal to the corporation, collaborative in outlook, and keen to keep a low profile. The outcome is impressive — and demonstrates that successful innovation can take many different forms. ■

"When oil prices went through the roof last year, it was Toyota and Honda — not General Motors or Ford — who were ready at the starting gate with 'hybrid' vehicles."

RESEARCH HIGHLIGHTS

CHEMISTRY

Express delivery

Org. Biomol. Chem. doi:10.1039/b504988a (2005)
It is difficult to design drugs that can cross the blood–brain barrier. But a team led by Jean Louis Kraus of the Mediterranean University in Marseilles, France, has found that ascorbic acid can act as a delivery vehicle.

Ascorbic acid is actively transported from the blood to the brain, where it has a protective function. The researchers showed that attaching ascorbic acid to a drug called DAPT, used to treat Alzheimer's disease, dramatically increased DAPT uptake into the brains of live mice. Furthermore, the drug's activity *in vitro* was increased by linking it to ascorbic acid.

EARTH SCIENCE

Taking the strain

Earth Planet. Sci. Lett. **234**, 421–435 (2005)
Monitoring barely perceptible tremors could reveal whether geological slip faults are likely to spawn larger earthquakes, propose Rocco Malservigi, now at the Ludwig Maximilian University of Munich, and his colleagues.

The researchers studied the Hayward fault, which is part of the San Andreas fault system in California. The build-up of strain along this slip fault is gently released as segments of rock creep past each other. But jammed regions have triggered significant earthquakes in the past. Modelling the 'microquakes' allowed the researchers to map the jammed areas far below the surface.

STEM CELLS

Cancer source

Cell **121**, 823–835 (2005)
The starting point for some lung cancers (pictured right) may be a newly discovered type of stem cell, according to Tyler Jacks's team at the Massachusetts Institute of Technology in Cambridge.

The team discovered a population of stem cells that produce some of the lung's specialized cells — both Clara and alveolar epithelium cells. Experiments in a mouse model for lung cancer implicated these stem cells in the formation of tumours. The *K-Ras* cancer gene that initiates tumours in mice was shown to drive proliferation of the stem cells, and tumours developed in regions where stem cells were found. Lung injuries that increased the incidence of tumours also boosted the number of stem cells, suggesting that these cells explain the link between tissue damage and lung cancer.

Bad vibrations

Anim. Behav. doi:10.1016/j.anbehav.2004.09.019 (2005)
Tree-frog embryos have a remarkable ability to hatch early when their eggs are attacked by snakes. This depends on sophisticated sensing of vibrational cues, reports Karen Warkentin of Boston University, Massachusetts.

Eggs of the red-eyed tree frog, *Agalychnis callidryas*, usually hatch after seven days, but the embryos can emerge up to 30% earlier to escape a predator's attack. Warkentin shows that eggs are more likely to hatch when exposed to vibrations recorded from a snake attack than when exposed to recordings of the vibrations caused by heavy rain. The embryos must therefore be able to distinguish between these different kinds of motion.



K. M. WARKENTIN

GENETICS

Muffled mice

Proc. Natl Acad. Sci. USA
doi:10.1073/pnas.0503739102 (2005)
The *Foxp2* gene has previously been associated with speech and language in people, and with song learning in zebra finches and canaries. By engineering mice with one or two disrupted copies of *Foxp2*, a group led by Joseph Buxbaum of the Mount Sinai School of Medicine in New York has now shown that disrupting even a single copy of the gene impedes the ability of baby mice to call out to their mother.

If confirmed, the link between *Foxp2* and

communication in mice could make the mouse brain a useful model for addressing questions about the control of speech and articulation in humans.

CELL BIOLOGY

Disentangling DNA

Nature Struct. Mol. Biol. doi: 10.1038/nsmb953 (2005)
Mutations in the *BRCA1* gene are associated with breast cancer, and it is well known that the protein produced by this gene is important for tumour suppression and the repair of damaged DNA.

Researchers led by Junjie Chen at the Mayo Clinic in Rochester have identified a new role for the BRCA1 protein in maintaining the integrity of DNA in normal cells. They show that the BRCA1 protein binds to and enhances the activity of the topoisomerase II enzyme that helps untangle DNA and segregate chromosomes when cells are replicating their DNA.

COSMOLOGY

Ripples in space

Phys. Rev. Lett. (in the press)
Preprint astro-ph/0412066 at <http://arxiv.org> (2004)
Microwave radiation left over from the Big Bang, known as the cosmic microwave background, contains ripples that reflect the

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

distribution of matter in the 300,000-year-old Universe.

Roberto Trotta from the University of Oxford and Alessandro Melchiorri of the University of Rome 'La Sapienza' claim they can detect fluctuations in this radiation caused by the distribution of neutrinos — weakly interacting particles that formed in the first minutes of the Universe. Trotta and Melchiorri base this claim on data collected by the space-based Wilkinson Microwave Anisotropy Probe during its first year in orbit. Their find seems consistent with the standard model of cosmology, but further data are needed to strengthen their assertion.

BACTERIOLOGY

Deadly mix up

J. Exp. Med. doi:10.1084/jem.20050112 (2005)
Cerebrospinal meningitis can strike swiftly and fatally, which is surprising because the bacteria that cause it, *Neisseria meningitidis*, normally reside benignly in the nose.

Researchers led by Xavier Nassif of the National Institute for Health and Medical Research in Paris have shown that a phage — a virus that invades bacteria — may participate in this apparently unpredictable switch to infectiousness.

The researchers compared the genomes of different strains taken from a large epidemiological collection in the Czech Republic. A particular gene cluster was detected in all the strains isolated from patients with disease, but in only 10% of the non-pathogenic strains. This cluster was identified as phage DNA.

BIOTECHNOLOGY

Muscle booster

Nature Biotechnol. doi: 10.1038/nbt1109 (2005)
Engineered muscle can now be grown *in vitro* with ready-made blood vessels, thanks to a technique developed by Robert Langer of the Massachusetts Institute of Technology and his colleagues. This should make it possible to grow thicker tissue samples for transplants, because the vessels supply nutrients to cells deep within the sample.

Langer's approach is surprisingly simple. His team mixed endothelial cells and precursors of mural cells — the cells that form blood vessels — into a culture of muscle precursor cells, known as myoblasts. The endothelial and mural cells self-assembled into a vascular network as the muscle tissue grew. When the muscle was transplanted into mice, the vessels grown *in vitro* hooked up with the animal's existing vessels, improving the implant's survival.

CANCER

Skin deep

Nature Genet. doi:10.1038/ng1586 (2005)

A mouse model of human skin cancer has been created by Paul Khavari's group at Stanford University, California, allowing the researchers to investigate the genetic mutations that underpin the disease.

Although many genes have been linked to tumours known as melanomas, it is not known which mutations have an active role in initiating tumour development. The team engineered human cells from the base of the skin's surface layer to express some of these mutations. They added these cells into human skin tissue grown on mice. The progress of the disease revealed the potency of the selected mutations, validating an approach that might now be applied to other human tumours.

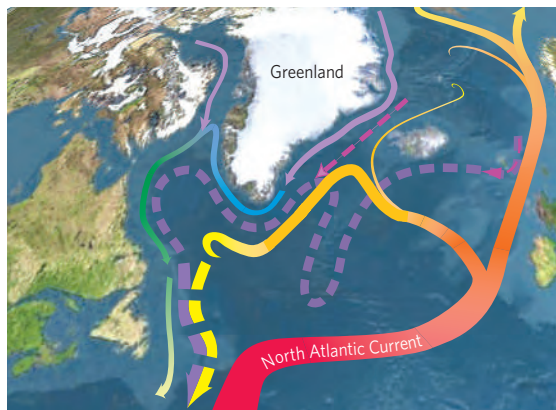
CLIMATE CHANGE

Watery world

Science 308, 1772-1774 (2005)

The volume of fresh water that has flooded into the North Atlantic Ocean since 1965 has been quantified for the first time, providing hints for how climate change may affect ocean currents (pictured).

Ruth Curry from the Woods Hole Oceanographic Institution in Massachusetts and Cecilie Mauritzen from Oslo's



Norwegian Meteorological Institute calculate that a torrent of fresh water caused a famous 'salinity anomaly' in the 1960s, but had little effect on ocean currents because it mostly ended up in remote subpolar basins. Since then, an average of 100 cubic kilometres of fresh water from rain, rivers and melting ice has diluted the upper layers of the Nordic seas each year, affecting the water density. Within a century, the seas' decreasing salinity could weaken the Atlantic circulation that takes cold, dense water southwards and brings vital, warm water northwards.

JOURNAL CLUB

Christopher Miller
Brandeis University,
Waltham, Massachusetts

A biochemist whose expertise lies in observing ion channels spots parallels in the study of protein synthesis.

"What walks on four legs at dawn, two at noon and three at sunset?" asked the Sphinx. Oedipus' correct answer was "Man, who crawls as an infant, strides as an adult, and hobbles with a stick in old age." But I would say that techniques in science follow the same trajectory.

I first witnessed this three decades ago, when electrophysiological recording made it possible to see single molecules in cells' ion channels. The method's morning was characterized by 'spot-the-blip' papers, whose results simply said: Look at me, I'm a protein! Only later, as noon approached, did the method allow us to explore mechanisms.

Now I am watching the exceptionally sexy techniques of single-molecule fluorescence travel the same path. An initial flurry of papers simply described using a light microscope to see single macromolecules. But researchers have started to address mechanistic questions, as in Scott Blanchard's paper last year (S. C. Blanchard *et al. Nature Struct. Mol. Biol.* 11, 1008-1014; 2004).

This paper attacks a central question concerning the birth of bonds in proteins. The authors track fluorescent versions of molecules called transfer RNAs, which carry the first amino acids of a protein chain into the cell's translation machinery. They clearly see the transfer RNAs going through two stages of checks before allowing the bond between the amino acids to form, some 100 milliseconds later. Such processes have never before been observed directly.

From my own ion-channel experience, I anticipate a long and mechanistically informative midday for the single-molecule fluorescence field, before the inevitable methodological decrepitude of evening sets in.

NEWS

Drug targeting: is race enough?

WASHINGTON DC

The expected approval in the United States of the first drug targeted to a specific racial group is sparking debate about the future of 'personalized' medicine.

Enthusiasts predict a future in which people are given genetic tests to help choose the drug to which they will respond best. But some experts worry about the precedent of accepting race as a crude marker for underlying biological differences — which could still leave many individuals being treated with drugs that don't work well for them.

Last week, an advisory committee to the Food and Drug Administration (FDA) recommended that BiDil, a drug for congestive heart failure, should be approved for sale with a label designating African Americans as the target population. The FDA usually follows the advice of its experts and is expected to make its decision by 23 June.

Francis Collins, director of the National Human Genome Research Institute in Bethesda, Maryland, says that the drug's effectiveness in blacks is "something to celebrate". Even so, he argues: "We should move without delay from blurry and potentially misleading surrogates for drug response, such as race, to the more specific causes."

BiDil is a combination of isosorbide dinitrate, used to treat angina, and hydralazine,

which lowers blood pressure. It is made by NitroMed of Lexington, Massachusetts. The drug was initially tested in a population that was about two-thirds white and one-third black by another company. The results weren't persuasive, and the FDA turned down a request for marketing approval in 1997.

But Jay Cohn, a cardiologist at the University of Minnesota, reanalysed the data and found that the black patients had responded much better to the drug.

Cohn and NitroMed were granted a precedent-setting patent on BiDil as a racially targeted drug, and the company launched a new trial in 1,050 black patients. BiDil reduced deaths by 43%, proving so successful that the trial was stopped early, in 2004.

The drug's anticipated approval has been greeted with enthusiasm by both cardiologists and the drugs industry. "This is the most important advance in the care of black people that we've seen in my lifetime," Charles Curry, president of the International Society on Hypertension in Blacks, told the FDA.

But Collins is wary of using the "biologically inaccurate and socially dangerous" surrogate of race, rather than pushing researchers and companies to investigate the genetic and environmental factors that determine individual

differences in drug response. Even if blacks respond better on average to BiDil than whites, he points out, the drug will still be ineffective for those who don't possess a particular cardiac physiology or combination of genes. There may also be a minority of whites who would benefit from taking BiDil.

Collins contrasts BiDil with Iressa (gefitinib), the AstraZeneca drug whose effectiveness at treating advanced lung cancer was so disappointing that the FDA last week restricted its use to current users and patients in clinical trials. Nevertheless, in about 10% of patients on Iressa, lung tumours shrink rapidly. Japanese patients are three times as likely as whites

to fall into this group. But the underlying difference is that patients who respond well have specific mutations in the receptor for epidermal growth factor.

This opens up the possibility of using Iressa to treat people whose tumours have these mutations. "Wouldn't it be unfortunate if at this point all we knew is that there is a better chance of responding if you are Japanese?" says Collins.

But will drug companies have sufficient incentive to go beyond using race or other crude surrogates, when to do so would entail

"We should move without delay from blurry surrogates for drug response to more specific causes."

Trouble brews over contested trend in hurricanes

The debate over whether global warming is making hurricanes worse has been nothing if not stormy.

The issue came to a head in January, when leading US meteorologist Chris Landsea resigned from the Intergovernmental Panel on Climate Change, complaining that a colleague on the panel, Kevin Trenberth, had supported a link between warming and storms in a press conference. Now, just in time for the 2005 hurricane season, Trenberth has clarified his views in print (*Science* **308**, 1753–1754; 2005). He argues that the intensity, if not the frequency, of hurricanes and typhoons will increase as the oceans warm.

The hurricane seasons from 1995 to 2004 have been far above the long-term average in terms of the number of storms and accompanying rainfall. However, most scientists are still



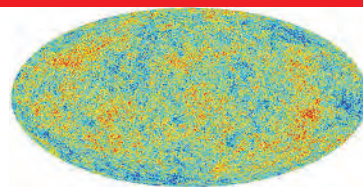
Open season: the latest hurricane analysis will wind up those who say that worsening storms are not related to global warming.

extremely cautious about connecting hurricane activity and global warming. Since satellite detection started 35 years ago, there has been no detectable trend in hurricane frequency, points out modeller Kerry Emanuel of the Massachusetts Institute of Technology in Cambridge, Massachusetts.

But Trenberth, head of the climate analysis

section of the National Center for Atmospheric Research in Boulder, Colorado, argues that because the number of hurricanes is relatively small, and fluctuates in cycles of various lengths, proving the existence of a trend from weather records is extremely difficult.

He has looked instead at how hurricanes form. "Trends in human-influenced environ-



NEUTRINO RIPPLES SPOTTED IN SPACE
Universal lumpiness is imprinted in mysterious particles.
www.nature.com/news

TROTTA/MELCHIORRI/OXFORD UNIV.



Positive discrimination: BiDil, a drug to prevent heart failure, will be targeted at black patients.

China's chicken farmers under fire for antiviral abuse

TOKYO

The much-feared H5N1 strain of bird flu has become resistant to one of the most effective antiviral drugs against it — and it seems that Chinese farmers' use of the compound in chickens is to blame.

This week the accusation was formally made — and formally denied. But at some point after 1997 the H5N1 strain became resistant to the amantadine family of antiviral drugs, and Chinese officials have now pledged to investigate the claim.

On 18 June, *The Washington Post* reported that Chinese farmers, encouraged by government officials, had been routinely using amantadine drugs in chickens. The United Nations Food and Agriculture Organization (FAO) and the World Health Organization (WHO) then questioned the Chinese government, which denied the reports, although officials did not comment on whether farmers were using the drugs.

The FAO's avian-flu surveillance network coordinator in China, Fusheng Guo, told *Nature* that the drugs have been widely used to combat the H9 family of viruses in chickens. Guo, who was a private consultant for Chinese farmers throughout the 1990s, says that he warned farmers not to use amantadine, because residues in the meat could make human viruses resistant. But, at the time, he did not worry about it in relation to H5N1. "Now I believe it's a serious problem," he says.

According to Guo, the agriculture ministry was probably not aware of the use. "The men selling it knew it was illegal," he adds.

China's health ministry told the WHO on 21 June that the problem is "worth following up in the closest way possible", according to Roy Wadia, the WHO spokesman in Beijing. Surveillance and testing will now be required to find out how widespread the use has been and how common resistance to amantadine has become.

The first H5N1 strain that infected humans, in Hong Kong in 1997, was sensitive to the amantadines. But resistance to them, discovered in 2003, has left another — more expensive — family of antivirals, including oseltamivir (sold as Tamiflu) and zanamivir (Relenza), as the only line of defence against the virus. This leaves the poor countries of southeast Asia without a low-cost option.

David Cyranoski

costly research and might narrow their target markets? Experts agree that they might not. But some point out that further studies could in some cases expand the potential market, and pinpoint the markers predisposing some people to dangerous side-effects.

"If I were a drug company executive, in

addition to finding out about what works, I might be able to find out what causes problems, and save myself some liability," says Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania in Philadelphia.

Meredith Wadman

mental changes are expected to affect hurricane intensity and rainfall," he concludes.

One simulation that Trenberth reviewed suggests that warming tropical oceans will stretch the upper limit of cyclones' potential strength.

"Most storms may actually not reach the limit," says Tom Knutson, a co-author of the simulation based at the US National Oceanic and Atmospheric Administration in Princeton, New Jersey. "But in principle, Trenberth's conclusions are consistent with our studies."

Trenberth also argues that higher sea surface temperatures in the Atlantic Ocean and increased water vapour in the lower atmosphere — caused by global warming — are to blame for the past decade's intense storms.

His conclusions will not please some in the meteorology community. In an upcoming paper in the *Bulletin of the American Meteorological Society*, Landsea, Emanuel and colleagues argue that there is no proven link between greenhouse-gas emissions and hurri-

cane behaviour. They point out that even if there were a trend that had been missed in weather records, the change would have to be quite small relative to the year-to-year variability that already exists.

Trenberth counters that sceptics are ignoring the evidence. "I am trying to get people to think about things in a different fashion," he says. "The point is that all meteorological events around the world are influenced in some way by global warming."

In any case, everyone is hoping that there will be fewer severe storms this summer than last — when four strong hurricanes struck Florida, and Japan was hit by a record ten typhoons. They also hope that the storm that came ashore unexpectedly in Brazil on 28 March 2004 will not be a harbinger of things to come: Catarina, as it was christened, was the first ever recorded hurricane to develop in the southern Atlantic Ocean.

Quirin Schiermeier

SPECIAL REPORT

Databases in peril

Life-sciences databases are in crisis, say their operators, as funders keen to support exciting new projects lose interest in maintaining existing services. *Nature* investigates the scale of the problem.

A lack of stable funding is threatening biology's core databases. Unless funding agencies set aside dedicated grants, the fear is that researchers will lose access to information vital to their work.

Several major international databases and research centres, including the European Bioinformatics Institute (EBI) at Hinxton near Cambridge, UK, face funding cuts. And the outlook for specialist databases is even worse: more than half of the operators contacted by *Nature* say their databases are updated sporadically or not at all because no funding was available after their original grants expired.

"There is a funding crisis right now," says Rolf Apweiler, a member of the EBI and head of the UniProt/Swiss-Prot protein-sequence database.

"It's a paradox," adds Lincoln Stein, a bioinformaticist at Cold Spring Harbour Laboratory in New York. "The funding system assumes that projects have a lifespan of three to five years. But if biological databases are to do their job, they need funding for a decade or so."

Life-sciences databases have proliferated over the past decade, driven by genome-sequencing efforts and easy Internet access.

For some scientists, resources such as UniProt are as much a part of the basic research infrastructure as reagents and test-tubes. The EBI's website, for example, recorded 2 million hits on a single day this April. Researchers use the site to access everything from molecular structures to nucleotide sequences. Hundreds of smaller databases, often maintained by individual labs, focus on molecules and genes associated with particular functions or species.

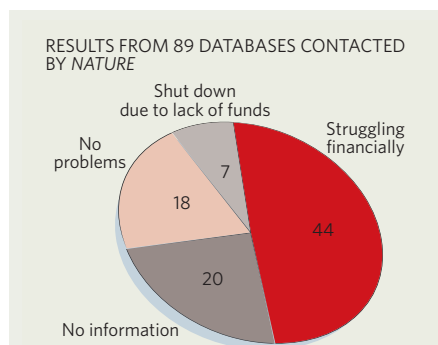
But as the number of databases mushrooms, many operators are finding that once their initial money has run out, funding agencies show little interest in helping maintain their service.

EBI director Janet Thornton is using the institute's reserves to support three databases whose funding has run out, including InterPro, an archive of data on protein families. "If we don't get new money we'll have to halve the number of staff on those projects," she says.

Thornton and others say the problem is particularly acute in Europe, where most grants are tied to original research. "Researchers feel like they have to invent new projects every three years to get money," says Thornton.

But databases in the United States are also feeling the pinch. The Alliance for Cellular Signaling, an ambitious ten-year attempt to amass data on the chemical signals inside cells, has scaled back its operations following a mid-project review. Funders at the National Institute of General Medical Sciences ruled last month that the project will receive less than half of the \$5 million a year it had asked for. The alliance says it will now have to shut five of the nine labs that are generating data from mouse-cell experiments.

Another key North American resource — the Biomolecular Interaction Network Database (BIND) — also faces cuts. Many journals, including *Nature*, routinely send their papers to BIND staff, who curate records on almost 180,000 molecular interactions (see page 1028). Last month BIND was forced to cut 33 jobs when a grant application to the Canadian government, which provided money to establish the database, fell through. Although BIND will continue to function thanks to money from the Singapore government, plans to integrate with other databases have been put on hold.



Survival of the fittest?

Nature contacted 89 databases listed in the Molecular Biology Database Collection (*Nucl. Acids Res.* **28**, 1-7; 2000) to see how many still have funding five years on. Of these, 51 reported that they are struggling financially. Seven of these have closed; the rest are being updated sporadically in their owners' spare time.

Not all these databases necessarily deserve to receive continued funding, especially as technology advances and competing resources are set up. But bioinformaticists are concerned by the sheer number that cannot find the money to provide an ongoing service. **Z.M.**

"Canada is good at starting up projects like this, but there is no mechanism for continuing them," says Chris Hogue, principal investigator at the Blueprint Initiative, the Toronto-based organization that runs BIND.

Quest for novelty

The cutback at the Alliance for Cellular Signaling may be the result of a conscious change of heart by funders, but bioinformaticists say other cuts are part of a broader problem facing databases: agencies want to fund innovative and hypothesis-driven initiatives, rather than ongoing infrastructure projects.

"Long-term maintenance is expensive," says Carol Bult of the Jackson Laboratory in Bar Harbor, Maine, home of the Mouse Genome Database. She says it costs around US\$4 million a year to run. The resource is widely used and Bult is confident that funding will be renewed this year, but many other databases aren't so lucky. "We've faced this issue for a decade, but the funding agencies haven't caught up."

Smaller, cheaper databases are in even more trouble. *Nature* contacted 89 databases operating in 2000, and more than half said they are now struggling financially. Seven databases have folded, and many others are updated on an irregular basis as a labour of love by their owners (see 'Survival of the fittest').

"It is far more difficult for an individual researcher to obtain funding to maintain a database than it is to initiate a new project, even though constant updates are very important for the value of the database," says Ikuo Uchiyama of the National Institute for Basic Biology in Okazaki, Japan. He has had to temporarily halt updates to his Microbial Genome Database owing to lack of funds.

Students often end up filling the gaps. Will Ray of Ohio State University began working on PACRAT, which pre-processes genome data from the GenBank database, when he was a graduate student. "Development of PACRAT was squeezed out of my doctoral adviser's grant," he says. "But the hardware and system support were, and still are, donated to the university out of my own pocket."

Many warn that if the situation continues, future research could be severely compromised. "Every working biologist relies on these resources," says Thornton of the EBI's databases. "Over the next 20 or 30 years, the whole



Key services: databases are as much a part of scientists' basic equipment as test tubes and reagents

of biology will be built on protein structure information available through these databases.”

The solution, say operators, is for funding organizations to set aside dedicated funding streams. Stein suggests that databases should be funded for longer than research projects, perhaps for five years, and that they should be judged separately. Bult adds that the system will need to be tailored so it can handle big projects such as the mouse genome database, as well as smaller, specialist archives that are used by just a few hundred researchers.

Any funding would also need to be vigorously peer-reviewed, and require databases to show they provide a service that is actually used. Bioinformaticists say that some databases lack community support and don't deserve continued funding. “You wouldn't want ten years of guaranteed support,” says Stein. “That would encourage waste and drift.”

Some agencies have already taken steps to provide earmarked funds. The Jackson Laboratory, for example, is applying to a panel of the National Human Genome Research Institute that focuses on large projects that need input from biologists and computer scientists. The Wellcome Trust, a London-based medical charity,

gives the EBI almost €2 million (US\$2.4 million) a year. And the next round of the European Union's Framework Programme will include a funding stream for infrastructure projects, including databases. This was originally set to distribute €3.5 billion, although that figure is likely to be cut substantially by the time the programme begins in 2007.

Operators say that a more general change in attitude is needed if biologists are to continue to enjoy the database services that they currently take for granted.

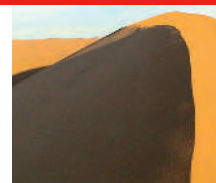
Apweiler suggests that databases should club together to make their voices heard. “They are small and not well-connected,” he says. “They need to unite and demand community resources, and then perhaps the funding bodies will respond.”

One success story cited by Apweiler is FlyBase, a resource partly funded by the US National Institutes of Health that links information on *Drosophila* from many databases.

But Apweiler warns that the money for the database comes at the expense of extra research funding in the field. “To increase these resources, the community needs to be willing to accept cuts in funding for individual projects,” he says. ■

Zeeya Merali and Jim Giles

**IMAGE
UNAVAILABLE
FOR
COPYRIGHT
REASONS**

**ARID PROSPECTS**

Expanding deserts mean many of the world's poorest people will have to seek new lifestyles.

www.nature.com/news

N. JONES

Earth holds comet smash in its sights

The largest coordinated observing campaign in astronomy's history is getting under way. More than 60 observatories — and virtually all of the world's big telescopes — are preparing to view the collision between NASA's Deep Impact spacecraft and the unsuspecting comet Tempel 1 on 4 July.

The copper-fortified craft, released from a flyby spacecraft carrying the primary imaging instruments, will smash into the 4×14 -kilometre target at 10 kilometres per second. This will gouge a hole and throw out material from the comet's interior that scientists hope will reveal much about its make-up, and thereby the origins of our Solar System.

Meanwhile, the flyby craft will record the impact from close up with cameras and spectrometers — but only for about 13 minutes. That leaves more distant telescopes on Earth and in space to watch what unfolds as dust, ice and gas continue to spew out from the comet for anywhere between hours and weeks. Earth-based instruments can also view the action at wavelengths unavailable to the spacecraft.

Ground telescopes have already made a

major contribution, says Karen Meech of the University of Hawaii, who coordinated the Deep Impact observing campaign. At the start of the mission in 1999, the diameter of the comet's nucleus was known only to within 50%. Years of viewing by Earth-based telescopes (see graphic) mean that it is now known to within a few hundred metres. And ground observers have been monitoring the comet's dust output, partly to gauge the hazard to the spacecraft, and also identifying chemicals in the surrounding coma, so that these can be compared with what flies out of the comet on impact day.

One of Meech's main headaches has been taking into account the different constraints of space- and Earth-based observatories. The Deep Impact team recently decided to push back the spacecraft's arrival time at Tempel 1 by 17 minutes so that the Hubble Space Telescope, whose exact orbit cannot be predicted months in advance, will be in the correct position at impact time. And early on, the team had to choose whether to optimize viewing in Hawaii or Chile, the world's two premier sites for astronomy.

Chile's telescopes had the advantage of being more spread out, which improved the chances of not being rained out on impact day. But timing the impact for viewing in Hawaii meant better coverage from the NASA antennas that communicate with the spacecraft, so Hawaii got the nod.

Meech herself will watch from a control centre on the peak of Mauna Kea in Hawaii that will link astronomers at around 30 global observatories by live video — the first time this has been done for an astronomical event. Other telescopes worldwide will tune in before, during and after the encounter.

Added to that will be hundreds of amateur astronomers, "many of whom are absolutely excellent observers", says Beech, and members of the general public, who should be able to see the brightening cloud of dust spraying from Tempel 1 with binoculars, perhaps even with the naked eye. Viewers from Arizona west to New Zealand will see the impact as it happens, although viewers in Chile won't see the comet until 16 hours later.

Tony Reichhardt

Telescopes around the world have paved the way for Deep Impact



Kitt Peak National Observatory (4 m, 2.1 m)

Flagstaff, Lowell Observatory (18 m)

Hawaii



Mauna Kea, Keck Observatory (10 m)

Paranal, Very Large Telescope (8 m)
Cerro Tololo Inter-American Observatory (4 m, 1.5 m)



Davis Mountains, McDonald Observatory (2.7 m)



La Palma, Telescopio Nazionale Galileo (3.6 m)

Canary Islands



La Silla, Atacama desert, Danish Telescope (1.5 m)
MPG/ESO Telescope (2.2 m)
NTT Telescope (3.6 m)



Mt Bohyun, Bohyunsan optical astronomy observatory (1.8 m)

Southeast Korea



Mitzpe Ramon, Negev desert, Wise Observatory (1.1 m)

More than 60 observatories will be watching when Deep Impact crashes into Tempel 1 on 4 July. But several have been monitoring the comet since 1999.

ARIZONA: M. HANNA/NOAO/AURA/NSF; TEXAS: M. HARRIS/MCDONALD OBS.; HAWAII: KECK OBS.

SE KOREA: R. ARLT/ASTROPHYS. INST. POTSDAM; CHILE: ESO; CANARY ISL.: G. TESSICINI/FUND. GALILEO GALILEI-INAIF

ON THE RECORD

“You have to have an immediate effort to reduce greenhouse gases. Anything else is a fig leaf and a joke.”

US Senator John McCain urges his colleagues to tackle global warming in a highly contentious bill on energy policy.

“We cannot evade a liberalization of research on embryonic stem cells.”

German Chancellor Gerhard Schröder argues that his country must relax its stem-cell policy in a speech at the University of Göttingen.

“This is a whitewash. They took all of our science and reversed it 180 degrees.”

In the *Los Angeles Times*, Erick Campbell protests that the government altered a report he helped to write on the impacts of grazing. The retired biologist says it was the cattle industry, not the environment, that benefited from the edits.

NUMBER CRUNCH

A UK survey, released this week, found that many children wouldn't take science at all if they didn't have to. Why?

79% of British schoolchildren think scientists are **clever**.

7% think scientists are **cool**.

Source: OCR exam board

SCORECARD

**Sting fever**

Japanese women are clamouring for the chance to spend a night with the jellyfish at Fujisawa's aquarium. The experience is said to relieve stress.

**Slick move**

Philip Cooney, the White House official who altered climate-change reports, returns to the oil industry with a new job at ExxonMobil.

**Speech for the stars**

NASA astronaut delivers the first-ever congressional testimony from space, showing lawmakers a surefire way to rise above partisan politics.

**Milking it**

We feel it in our bones: sales of calcium tablets are set to soar. A ten-year study has shown that the mineral can prevent premenstrual syndrome, although no one knows why.



Currents of dissent: critics say that farms in deep water could create 'dead zones'.

NOAA

Bill on deep-sea fish farms brings wave of disapproval

SAN DIEGO

Scientists and activists have criticized proposed legislation that would push US fish farms into deep waters, beyond the reach of states' environmental controls.

The bill, introduced on 7 June in the Senate, would allow aquaculture pens between 5.5 and 370 kilometres off the US coast. Federal laws, not state regulations, prevail in this 'exclusive economic zone' (EEZ). The United States has the largest EEZ in the world, encompassing about 9 million square kilometres.

Tuna, salmon, halibut and cod could be farmed in the EEZ. The National Oceanic and Atmospheric Administration (NOAA), which drew up the bill, argues that it would create an industry to produce healthy food in an environmentally friendly manner. But others say that deep-sea fish farms could transmit diseases to wild fish and pollute the waters.

Fisheries officials foresee annual fish production worth \$5 billion. But this could produce as much nitrogen as 10 million pigs on land, according to the advocacy group Environmental Defense, based in New York City.

Nitrogen pollution can create 'dead zones' where few aquatic organisms live. "This can change the biology of the ocean," says Roz Naylor, an economist at Stanford University, California, who studies aquaculture.

If the bill becomes law, it could also lead to a stand-off between state and federal

authorities. Governors would be able to veto aquaculture in waters next to their state. Alaska's governor, Frank Murkowski, has already asked for a five-year moratorium so that more research can be done on the environmental and socio-economic effects of aquaculture.

Environmentalists advocate more federal controls similar to the state ones proposed in a bill now going through the Californian legislature. The bill, sponsored by state senator Joe Simitian of Palo Alto, would set environmental standards for fish farms in state-controlled waters.

Many observers believe such guidelines are needed. Peter Douglas, executive director of the California Coastal Commission, served on a NOAA scientific advisory panel several years ago that recommended environmental guidelines for fish farms in the EEZ. "As far as I can tell, they blew those off," says Douglas.

NOAA officials say they will address the issue of environmental standards during the upcoming discussions, and point out that they have done extensive background work. "This bill is ten years in the making," says Michael Rubino, an economist with NOAA.

The federal bill is being sponsored in the Senate by Democrat Daniel Inouye of Hawaii, where limited fish farming is under way in state waters, and Republican Ted Stevens of Alaska, a state that prohibits aquaculture altogether.

Rex Dalton

**FERTILITY CONFERENCE**

Highlights from the European Society of Human Reproduction and Embryology meeting.

www.nature.com/news

Retracted papers damage work on DNA repair

The retraction of a paper in the journal *Science* has left biologists picking up the pieces as revelations continue about the misdeeds of Tony Leadon, formerly a professor at the University of North Carolina at Chapel Hill.

The retraction, issued on 17 June, concerned a 1997 paper on a genetic disorder called Cockayne syndrome (P. K. Cooper, T. Nospikel, S. G. Clarkson and S. A. Leadon, *Science* 275, 990–993; 1997). Sufferers cannot repair their DNA properly, and the paper's authors reported that they had traced the syndrome's genetic cause to defective repair of oxidative damage — the sort of DNA injury caused by sunlight. But last week, three of the authors said that two figures contained invalid data, and that the paper's conclusions could no longer be supported. The fourth author, Leadon, refused to sign the retraction.

In 2003, a university panel found Leadon guilty of fabricating and falsifying findings in his research on DNA repair. Leadon left the

University of North Carolina on 31 March, after the verdict. He also left biologists struggling to sort out fact from fiction.

Two of Leadon's other papers were subsequently retracted. One of them claimed that a gene implicated in cancer, *BRCA1*, was involved in the repair of oxidative damage. But Philip Hanawalt of Stanford University in California — a leading figure in the field of DNA repair — says no other evidence has emerged to support the paper's conclusion. "His work on *BRCA1*, as far as I know, was simply fabrication," Hanawalt says.

The panel has made no public statement about the paper on Cockayne syndrome, but collaborators have been unable to replicate the results that it contains. There is some indirect evidence that problems with oxidative repair lie at the root of the disease, says one of Leadon's former collaborators, Priscilla Cooper of the Lawrence Berkeley National Laboratory in California. But, she

says, it will take time to establish the truth.

"The subfield that is working on oxidative damage is really having to redo everything from scratch," Cooper says. "This was quite a blow to me, and his other co-authors."

No one is quite sure what caused Tony Leadon to report false data or whether it was intentional; he has not responded to Cooper and Hanawalt's attempts to communicate with him, and he did not reply to *Nature's* attempts to contact him. But Hanawalt speculates that Leadon was driven to fabricate results after a test for DNA repair simply stopped working.

As the field tries to sort itself out, Hanawalt notes that there are other lasting impacts of the retractions. One of Leadon's co-authors, for instance, is searching for a faculty job, and must explain the situation to potential employers. "They can say, 'It wasn't my fault,' but the damage has been done," Hanawalt says. "This kind of thing can be devastating." ■

Erika Check

Japan steps up training of science communicators

Three of Japan's top universities are setting up the country's first programmes in science communication. The move is part of the government's efforts to keep young people in the sciences, says a spokesman for the Japanese education ministry. "We need more professionals who can explain to the public why researchers need taxpayers' money to continue their work," he says.

The universities of Tokyo, Waseda and Hokkaido will launch two-year graduate courses in science communication in the next two years. Content will vary from school to school, but could include journalism, teaching and public-relations skills for working at research institutes, funding agencies or museums.

To fund the courses, each university will receive up to ¥100 million (US\$920,000) each year for the next five years.

Antarctic research nations decide polluters must pay

Those responsible for polluting Antarctic waters or causing other environmental emergencies on the continent will now

be legally bound to pay for the clean-up operation.

At a meeting last week in Stockholm, the 28 countries with serious research interests in Antarctica adopted new rules that would hold individual operators or nations responsible for taking immediate action after an environmental disaster. If no action is taken, the responsible party would have to compensate other groups for their clean-up efforts or pay for future operations.

The pact extends the 1959 Antarctic Treaty, which protects the continent from military operations and promotes scientific cooperation between countries. Before the new agreement takes effect, each of the 28 countries must ratify the rules within its own legislature.

China and US streamline visas for visiting scientists

The US and Chinese governments have reached an agreement that will allow scientists and graduate students to move more easily between the two countries.

Historically, Chinese citizens have been one of the largest groups of foreign researchers working in the United States. But after the terrorist attacks of 11 September 2001, many found themselves undergoing

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Asking for access: Chinese scientists and students have protested over US visa problems.

lengthy security checks to gain visas. Delays were exacerbated by a bilateral agreement requiring Chinese students and scholars to reapply for a visa after entering the country twice or after six months, whichever comes first (*Nature* 427, 190–195; 2004).

Under the new deal, which came into effect on 20 June, Chinese nationals applying to study or carry out research in the United States will be issued with one-year, multiple-entry visas. This will make it easier for Chinese scientists working in the United States to travel to meetings abroad, and allow students to return home for holidays or family crises. Similar provisions will be granted to US scholars working in China.

Truce declared over free access chemical database

Tensions have eased between the world's largest science agency and the world's largest scientific society. The US National Institutes of Health (NIH) and the American Chemical Society (ACS) have suspended their public sparring over the NIH's free PubChem database, which the ACS says threatens its much larger, money-making chemical database.

A congressional report that accompanied the NIH funding bill for next year was expected to contain strong wording that would limit PubChem to including only molecules produced by NIH screening centres. But when presented to the House Appropriations Committee on 16 June, the report merely said that the NIH should "work with public sector providers to avoid unnecessary duplication and competition".

Both sides said they were happy with the result. NIH chief Elias Zerhouni reportedly worked hard to soothe ruffled feelings during a meeting between top officials on 3 June.

For the moment PubChem is not much more than a prototype, but it is expected to swell with data from new NIH screening

Polar researchers wheel out prototype snowmobile

Antarctic researchers could soon be travelling more safely over the snow and ice thanks to a concept vehicle unveiled on 20 June.

Created by London-based designer James Moon in conjunction with the British Antarctic Survey (BAS), the two-person vehicle comes with a small linked pod. The pod scouts a path about 30 metres ahead, following a route set by global-positioning satellites and using ground-penetrating radar to check for hidden crevasses.

The BAS is considering whether to develop the idea for the vehicle, dubbed Ninety Degrees South, into a fully functional model.



J. MOON

centres, nine of which were announced on 15 June.

NIH set for 'disappointing' increase in 2006 budget

A House of Representatives committee has approved a budget increase of \$142 million — a rise of just 0.5% — for the US National Institutes of Health (NIH) next year.

The increase approaches that requested by President George W. Bush in February. The Senate committee that sets NIH budgets has

not yet passed its version of legislation to fund the agency, but it traditionally boosts NIH funding above House levels. Before the budget becomes law, the two versions of the budget must be reconciled.

Supporters of biomedical research funding are disappointed with the committee-approved 2006 NIH increases, saying that they fall short of the projected rise in costs.

In a separate vote by the entire House, the National Science Foundation won a 3.1% increase to \$5.64 billion — \$38 million above the president's budget request.



A. HAAG

A trip of a lifetime

Are research expeditions to far-flung destinations as glamorous as they sound?

Amanda Haag joins a few research novices who gave up their holidays for science.

It is a cold, overcast May morning on a tiny island in the Gulf of Maine. Bill Clark and Karen Abbey are scrambling across algae-coated rocks, choosing their steps carefully to avoid seabird nests and poison ivy. One wrong move could send a protective gull parent into alarm mode, leaving the pair being screeched at, pooped on or, worse still, dive bombed.

Clark and Abbey have put themselves in this precarious position to measure eggs and to place chicken-wire cages over empty nests ready to catch a gull when it returns. When they secure a bird, it's over to Julie Ellis, an ecologist from Cornell University, Ithaca, New York, who coaxes the gull into a bag. With the gull tucked safely inside, the team takes blood samples and wing measurements, and fixes numbered identification tags to one of its legs.

I ask Clark if he has been dive-bombed yet this week. "I haven't been hit, other than this splotch here," he says, motioning to the smattering of well-placed gull dung on his back. "This is the adventure component of the trapping," quips Ellis. It might be an adventure, but for Clark this isn't his usual occupation —

remarkably, he has volunteered his time and money to be here.

Clark, 66, of Kresgeville, Pennsylvania, and Abbey, 39, of Harpswell, Maine, signed up to help Ellis with her research through a non-profit organization called the Earthwatch Institute in Maynard, Massachusetts. Clark is a retired school teacher, and the gull project is his second Earthwatch expedition. This makes him typical of most Earthwatch recruits, a third of whom are repeat volunteers. Abbey is also a high-school maths and science teacher and like many first-time volunteers has never done research before. She is using her trip as part of a programme to bring science to her students.

Public service

Earthwatch was founded by educators and scientists in Boston in 1971 as an alternative way to fund field-based scientific research. Since then, it has grown into an international organization that each year sponsors research to

the tune of some \$4 million, supporting more than 140 projects across 47 countries. It is a pioneer of what is now called 'citizen science' — when scientists involve members of the public in collecting their data. Every year, around 4,200 people from 80 countries volunteer for Earthwatch projects

— and the scheme has helped to generate thousands of peer-reviewed publications over the years. A similar organization called Biosphere Expeditions, based in Britain, started up in 1999, and smaller programmes can be

found at some individual institutions, such as the University of California, Davis.

For researchers, citizen science is a way to promote public outreach while getting help with labour-intensive data collection. For the volunteers, each expedition, lasting one to three weeks, is an opportunity to contribute to meaningful research in interesting locations. Some volunteers attend the same expedition multiple times: the current record-holder has notched up 20 trips to South Dakota to exca-

"The beach holds no attraction for me. I like to be with people who are thinking and learning." — Bill Clark

K. ABBEY vate the remains of mammoths. Clark is already planning his next Earthwatch expedition, this time with his wife, to Peru to look at the effects of ecotourism on macaws. Ordinary holidays have only limited appeal, he explains: “The beach holds no attraction for me. I like to be with people who are thinking and learning.”

The research that Clark and Abbey are helping with, at the Shoals Marine Laboratory on Appledore Island, involves studying the impact of gull populations on coastal ecosystems. The number of gulls in New England fell dramatically during the 1800s because their showy plumage was popular for adorning womens’ hats and gowns, and seabird nesting colonies were raided for eggs. But since all seabird species became protected in the 1900s, gull numbers have risen to such an extent that they now have the potential to significantly change their surrounding ecosystem.

Watch and learn

For Abbey, the expedition is a way of bringing maths and science to life for her students. She has designed a website with a daily journal that her students can access while she is on the island. She also got her statistics class to draw up a model of gull population growth, and she says that she hopes to do more fieldwork with her students in the future.

Earthwatch projects cover a wide range of interests and locations. A quarter of them fall into the social sciences, including ethnobotany, archaeology and public health. The rest incorporate some aspect of conservation, sustainable development or environmental monitoring; but all involve basic research and education. Projects run the gamut, from counting butterflies in Vietnam to measuring methane from termite mounds in Namibia.

Because Earthwatch sponsors expeditions all over the planet, the uninitiated might be inclined to think of the trips as holidays. But the expedition briefings do not mince words



Duck and cover: Karen Abbey uses a plastic bag to protect herself from gull attacks.

about what will be expected of participants, who may have to pay between \$500 and \$3,500, depending on the location and the length of stay.

The web page for the Maine gull project carries a clear warning: “Volunteers should be stout of heart for deflecting gull attacks and moderately agile for clambering over slippery, algae-covered boulders.” Blue Magruder, director of public affairs for Earthwatch, says that the organization is careful to encourage the ‘right’ volunteers — while deterring about half the people who make enquiries. “That’s good — that’s what you want to do,” says Magruder. “Because it’s no fun if somebody gets there and they’re bored after three days.”

Expedition scientists and volunteers agree that the trips don’t qualify as ecotourism, adventure tourism — or any other kind of tourism. “You don’t look at it as a vacation because you’re going there to support the

research,” says Clark. “Any vacation you get is very incidental.” Some volunteers say that if they wanted to tour an area, they wouldn’t choose an Earthwatch trip. Magruder recalls greeting one volunteer on his return: “I said, ‘How was Ireland?’ and he said, ‘Well, my two-metre square was fascinating.’”

Projects that are good candidates for citizen science usually involve labour-intensive or quantitative tasks, such as counting numbers of nests or logging data on zebra behaviour. In addition, the tasks must be fairly easy to learn. As a result, most expeditions tend to be field-oriented: working behind the bench in a molecular-biology lab wouldn’t lend itself to recruiting many volunteers, or to data that can be collected reliably with little training.

Some expeditions, such as Rolf Peterson’s moose and wolves project in Isle Royale, Michigan, are more physically demanding than they are scientifically intense. “The main thing is to get all those bones back,” says Peterson, an ecologist at Michigan Technological University, who leads expeditions to collect the remains of moose killed by the wolves that share the island. Peterson’s research involves backpacking for 16 kilometres a day through the backcountry with 2–4 kilograms of moose bones strapped to the volunteers’ backpacks, on top of a week’s worth of food and equipment.

Skilful display

Scientists from different projects agree that volunteers often bring with them unexpected skills. On this trip, Abbey, who has a background in computer programming, designed a database entry-form for Ellis to help her extract data and analyse them more easily. Ellis notes that as a lead scientist you have to be flexible in using the volunteers’ skills. Last summer, she had one couple who were 75 and 88 years old on her team. “I was a little nervous about it,” recalls Ellis. Rock climbing was clearly out of the question, but the woman was a retired mathematician and helped to devise new ways to organize Ellis’s data, and the man, a retired zoologist, was a keen observer of gull behaviour.

Whatever the task, the scientists give careful consideration to quality control. Ellis remembers one volunteer whose data on gull behaviour she later threw out because “the minute I turned around, he got on his cell phone”. Eric Brown, a marine ecologist with the National Park Service in Kalaupapa, Hawaii, who has led Earthwatch expeditions to coral reefs, has compared the data collected by his staff with those from volunteers. His team was trying to determine the percentage decline of fish species between pristine and disturbed reefs.

Brown found that, although volunteers identified fewer fish than professional staff, the percentage decline the volunteers recorded was comparable to that noted by staff. But Brown’s study did illustrate the need to con-



Caught in a trap: a gull is snared pending measurements and tagging.

Underwater odyssey

Mark Patterson, now an associate professor at the Virginia Institute of Marine Science at the College of William and Mary in Gloucester Point, grew up on Jacques Cousteau documentaries. But he got hooked on coral-reef science when he won a high-school scholarship to take part in an Earthwatch expedition — at that time known as Educational Expeditions International — to the Bahamas in 1974.

Patterson describes his time in the Bahamas as magical. "It was great because we were diving all day long. It was a dream come true for me," he says. Patterson recalls Rick Chesher, the lead scientist on the expedition, being a good mentor: "He understood the aims of lay-person-driven science very well, and did an excellent job leading all these bankers, doctors and dentists." He also found time to talk to Patterson about what life in science was like. "I knew I wanted to be a marine biologist from the age of six. Now I knew I wanted to go to college and

become a coral-reef biologist," Patterson says. He went on to complete his undergraduate, master degrees and PhD at Harvard University and became an expert in coral-reef biomechanics.

Patterson also credits Chesher with encouraging him to pursue electronics as a means of making underwater instruments. Patterson first got into electronics when he was caught destroying metal-shop tools with a welder in seventh grade. As punishment, his teacher made him join the hand-radio club. But Patterson loved learning Morse code and got a licence for hand-radio operators, inspiring a lifelong interest in electronics.

"Chesher was talking to me about it at sea and said: 'You should really keep that up and not just concentrate solely on biology... you'll be able to do something that most marine biologists can't'," recalls Patterson.

In the early 1990s, Patterson was working with underwater robots and decided he wanted to build an autonomous underwater vehicle



D. KESLING

Mark Patterson shows off the underwater vehicle he helped to design.

(AUV). He tried to get funding from the Navy, from the National Oceanic and Atmospheric Administration and from the National Science Foundation, but he says that because he was not a certified engineer, "they told me to go away". So he and a family friend, Jim Sias, a world-class industrial designer, put together their own AUV, dubbed Fetch, on Patterson's kitchen table. They founded a company in 1996 — Sias Patterson

— to commercialize their product.

Patterson is currently using AUVs to study oxygen dynamics over coral reefs and is designing underwater robots with customized sonar that can classify different species of fish. He has now notched up a total of 84 days living underwater, which he says, "would have never happened without having this formative exposure under Earthwatch back when I was in high school". **A.H.**

sider which tasks are amenable to volunteers and which require a professionally trained eye. Earthwatch designed its own project to evaluate the contribution of volunteers, and found that although volunteers made some recording errors during fieldwork, similar errors were made by experienced scientists (J. Foster-Smith and S. M. Evans *Biol. Conserv.* **113**, 199–213; 2003).

Marie Studer, chief science officer for Earthwatch, says that the organization puts a tremendous amount of energy into helping scientists find constructive ways to integrate volunteers in their research. "We don't just accept or reject proposals," says Studer. "We see a good proposal that has value and potential, and we work very hard with the scientists to make it an Earthwatch project." The proposal process is much like any other funding request — it involves peer-review, and on average takes about a year, says Studer.

The scientists say that although most volunteers leave the excursions ready to return to their day jobs, 'conversions' do take place in the field (see 'Underwater odyssey', above). Andrew Russell, a glaciologist from the University of Newcastle Upon Tyne, UK, has led 25 Earthwatch expeditions to study glaciers in Iceland and Alaska, and knows of three volun-

teers who chose to pursue doctoral studies in Earth sciences. One volunteer who was on an Iceland expedition in 2002 is beginning his PhD work with Russell this autumn.

Earthwatch scientists always acknowledge the contributions of their volunteers in papers, posters and talks, although stewardship of the data remains solely with the lead investigator. But in the first summer of the Maine project, one of the volunteers decided to stay on for three more weeks to explore an interesting angle of the research. She later decided to pursue a graduate degree in ecology and will be a co-author with Ellis and her Cornell colleague Myra Shulman on an upcoming paper resulting from the work.

The degree of satisfaction among volunteers seems to be high. In a given year, probably only two or three — less than 0.1% — will leave the field early. And despite the rigours of data collection, few complain about being overworked, says Magruder. In fact, the most common complaints Earthwatch gets are from people who feel they haven't been worked hard enough or that their skills haven't been put to good use, she says.

But there is no shortage of work for the volunteers in Maine. For Clark and Abbey, Sunday is their day off, and brunch isn't until ten, but they are up and out the doors by 7:30,

determined to help tag as many gulls as possible before they eat. By mid-morning, it's raining relentlessly, so they return to base and take the opportunity to catch up on entering their data into the computer. "Julie told us we could take breaks whenever we wanted to, but we chose to just keep working," says Abbey.

It is clear that no one here is thinking of going home early. By the end of week one, Abbey and Clark were chatting with Ellis about her research with ease and authority. Both had become conversant in the intricacies of gull behaviour. "If she can get him to regurgitate something and she eats it, they'll probably mate," Abbey explains to me as we watch a gull pair from a distance. "That's their little dating ritual."

As Abbey's last day arrives, I'm preparing to leave too. Clark is staying on for a second week, and will soon be joined by a fresh team of volunteers. It is another bone-chilling, rainy New England morning, and after three nights in an unheated dorm, I'm ready to leave. I ask Abbey if she is pleased to be heading off — she's had only one shower all week and has been working until 10:00 every night. "I don't really want to go home — I wish I could stay," Abbey says, adding that once she started getting into the research, it was hard to walk away from it. "You just sort of take ownership of it," she smiles.

Amanda Haag is a freelance writer based in Boulder, Colorado.

"You don't look at it as a vacation because you're going there to support the research. Any vacation you get is very incidental." — Bill Clark

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

BACK TO OUR ROOTS

It was cold and clammy, but it changed the rules of life for ever.

Helen Pilcher goes in search of the ancestor of all animals.

Some geneticists have all the luck. While most are slavishly chained to the bench pipetting liquid, Werner Müller from the University of Mainz, Germany, gets to ponder the origins of life as he dives for sponges in the Adriatic Sea. His favourite spot is a 30-metre-long cave off the coast of Croatia, where sponges at the grotto's entrance are bright yellow, a hue bequeathed to them by the bacteria they contain. Sunlight pours through a hole in the ceiling. "It's a beautiful place," he says.

Müller hopes that his bounty will shed light on one life's biggest mysteries: how did animals come into being? Their birth is still shrouded by the mists of time, but scientists do know something special happened in the ocean around 600 million years ago when a group of single-celled creatures joined forces to form the first ever animal body, or 'metazoan'. Clubbing together allowed the cells to share the labour of living and paved the way for the evolution of specialized cell types, such as muscle, nerve and stem cells. This ancestor of all animals, known as the urmetazoan, would have needed a genetic blueprint for its structure or body plan. And this plan was the raw material that evolution acted on to give us the dazzling variety of animals we see today (see key examples in Graphic, opposite).

So it's no surprise that Müller and others want to find out more about this extinct animal. They want to discover how single cells took the pivotal step to make that first body, why animals evolved the way they did, and what it was about that early environment that kick-started animal life.

But with the urmetazoan dead for more than half a billion years, studying it is something

From the family album: sponges are probably the closest living relative to the first multicelled animal.

of a challenge. No fossils have been found, so there are no physical clues to its appearance. As a result, scientists are studying its closest living relatives: sea sponges and single-celled animals called choanoflagellates. By comparing these with each other — and with more complex animals such as mammals — they hope to build up a picture of the urmetazoan's genetic make-up and physical characteristics.

Simple blobs

Researchers have long regarded sponges as the most primitive form of animal life. As such, they are likely to be the most similar to the urmetazoan. At first glance, sponges seem simple. They have no gut, no brain, no obvious front or back, left or right. Adults pump water through a system of canals and cavities to extract food. "But they're much more sophisticated than the amorphous blob you see in the bath tub," says Müller.

For example, sponges are made of many different cell types. They have collar cells that beat their whip-like tails to create a water current inside the sponge's body, drawing food in and washing waste away. Stem cells give rise to sperm and egg, and epithelial 'skin' cells provide a protective barrier from the outside world^{1,2}. Cellular variety like this is the essence of multicellular life. And, just like in sponges, specialized or 'differentiated' cells would have allowed the urmetazoan to feed and reproduce simultaneously, making it more efficient than its unicellular neighbours, who could do only one job at a time.

Genetic comparisons between sponges and younger members of the animal tree provide

more insight into how these cells all worked together. Sponges, for example, have proteins called integrins on the surfaces of their cells. These tether the cells in place by sticking to another protein called collagen, which surrounds the cells³. Such proteins are found in more complex metazoans, so it is likely that the urmetazoan used a similar cellular adhesive.

But the cells also have to communicate. They have to organize themselves into a body with multiple cell types, so strategies are needed to tell the cells where and what they should be. Like more complex animals, sponges solve this problem by using specific molecules to guide differentiation and migration as the cells develop in their embryos.

Animal magnetism

Six hundred million years ago, singletons had the Earth largely to themselves. Its oceans teemed with single-celled organisms such as choanoflagellates and bacteria.

But some 10 million to 50 million years later, the fossil record documents a series of early 'experiments' in multicellularity. The Doushantuo Formation in southern China contains fossils of dividing embryos. A mishmash of strange animal-like creatures known as the Ediacaran fauna (such as the *Dickinsonia* pictured right) made an appearance.

What prompted single cells to get together? Geological records suggest that global oxygen levels rose dramatically about 600 million years ago¹⁰. The new 'breathable' concentration allowed oxygen to diffuse across

Bernard Degnan, a geneticist at the University of Queensland, Australia, is studying embryos of a sponge called *Reniera* to understand more about how body plans emerged in evolution. *Reniera* embryos contain at least 11 specialized cell types arranged in a particular pattern⁴. Pigmented cells, for example, are initially found on the embryo's surface, but then migrate to one end where they form a dark spot and then a ring.

Degnan speculates that cell migration and differentiation in these embryos is controlled in part by soluble chemicals that diffuse along the embryo's body creating a concentration gradient, a system also used in higher animals. Müller has isolated one such chemical from another sponge, *Suberites domuncula*⁵, and thinks that it may influence cell differentiation in both embryonic and adult sponges.

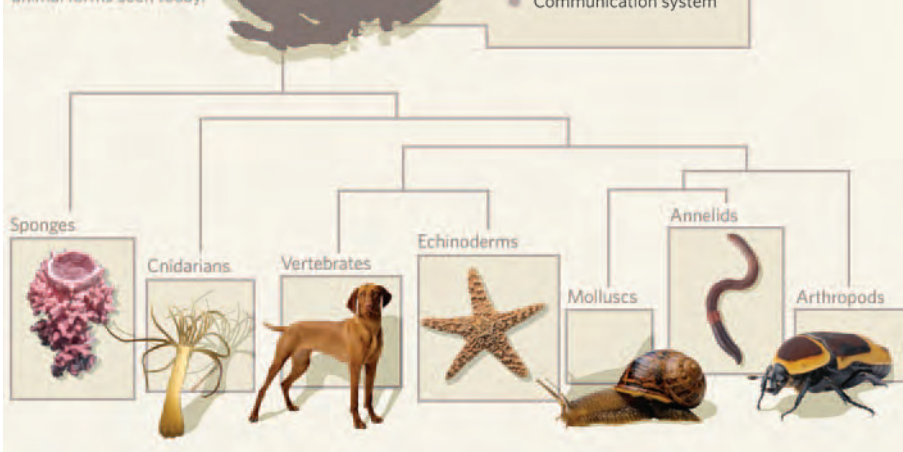
There are signs that many other molecules associated with development in animals also occur in sponges. The Wnt family of proteins, for example, influences how cells become specialized and also helps to lay down the key spatial coordinates of the body plan in complex animals. Sponge cells make the Frizzled protein, a receptor that is activated by Wnt proteins⁶. And they also make a variety of metazoan-like transcription factors — proteins involved in controlling gene expression — that are key players in development^{7,8}.

The fact that these genes occur during development in all existing animal lineages hints that they were playing a regulatory role in the embryos of the first metazoan. "The urmetazoan was probably quite sophisticated in a developmental and genomic sense," says Degnan. This suggests that it already had the genetic toolkit to direct a body plan containing multiple cell types.

To find out where this toolkit came from, biologists are looking even further back in time, at the single-celled ancestors of the urmetazoan. Their modern-day descendants are choanoflagellates, unicellular creatures that look uncannily like sponge collar cells. Surprisingly, choanoflagellates harbour many

All toolled up

The animal ancestor, or urmetazoan, had a toolkit of genes that evolved to give the diversity of animal forms seen today.



of the tools needed for multicellular living.

Geneticist Nicole King from the University of California, Berkeley, has discovered that choanoflagellates express genes involved in cell adhesion and communication in animals⁹. "This tells us that the molecular machinery for multicellularity was on site before the transition to multicellularity took place," says King.

Parental traits

Like choanoflagellates, the urmetazoan's single-celled ancestors may have used signals relayed by proteins called tyrosine kinases to sense changes in the outside world. Cell adhesion might have helped unicellular organisms to form simple colonies — colonies that may have been an intermediate step between single cells and true multicellularity. The urmetazoan may then have recruited these genes for new purposes. "Evolution is an extremely dynamic system and paradoxically a very lazy one," says palaeobiologist Simon Conway Morris at the University of Cambridge, UK, who studies the origins of meta-

zoans. "It will co-opt whatever it can."

But this raises a puzzling question. If the toolkit was already there, why didn't animals evolve sooner? The answer seems to be that the catalyst for multicellular animal life may not have been genetic but environmental — in the form of rising global oxygen levels (see 'Animal magnetism', below).

Whatever the trigger, researchers hope that the completion of the choanoflagellate genome sequence, expected later this year, will yield fresh insights into the biology of this momentous transition. Genome sequences are also expected for the sea anemone *Nematostella vectensis* and the simplest known living animal, *Trichoplax adhaerens*. *Trichoplax* is just two cells thick, has only four cell types and looks like a giant multicellular amoeba. Its genome promises to be the smallest of any animal yet measured and should define the minimum set of genes needed for animal life. Sea anemones, which have a more advanced body plan, could yield information on the genetic mechanisms underlying body-plan formation.

So it seems that the urmetazoan was a sophisticated creature. But don't expect a photo reconstruction for the animal family album just yet. "At present, its size and shape are pure speculation," says Degnan.

Helen Pilcher is a freelance science writer based in London.



S. CONWAY MORRIS

multiple cell layers, giving animals the freedom to become more complex, says biogeochemist John Hayes from Woods Hole Oceanographic Institution, Massachusetts.

No one knows exactly what triggered the rise. Most researchers speculate that it happened when large amounts of organic matter were buried at the bottom of the ocean. This permanently removed carbon from the carbon cycle, upsetting the balance of carbon and oxygen in the water and causing the oxygen concentration to rise. Some suggest that the faecal pellets of tiny single-celled organisms may have caused organic matter in the water to sink. Another theory proposes that the change in oxygen levels was triggered by massive dust storms or the break up of the landmasses or 'supercontinents' that then dominated the globe.

H.P.

1. Müller, W. E. G. *et al.* *Int. Rev. Cytol.* **235**, 53–92 (2004).
2. Adell, T. *et al.* *J. Mol. Evol.* **59**, 41–50 (2004).
3. Brower, D. L., Brower, S. M., Hayward, D. C. & Ball, E. E. *Proc. Natl Acad. Sci. USA* **94**, 9182–9187 (1997).
4. Degnan, B. M., Leys, S. P. & Larroux, C. *Integr. Comp. Bio.* **45**, 335–341 (2005).
5. Schröder, H. C. *et al.* *FASEB J.* **14**, 2022–2031 (2000).
6. Adell, T., Nefkens, I. & Müller, W. E. G. *FEBS Lett.* **554**, 363–368 (2003).
7. Wiens, M. *et al.* *J. Mol. Evol.* **57**, S60–S75 (2003).
8. Adell, T. & Müller, W. E. G. *Biol. Cell* doi:10.1042/BC20040135 (2005).
9. King, N., Hittinger, C. T. & Carroll, S. B. *Science* **301**, 361–363 (2003).
10. Des Marais, D. J., Strauss, H., Summons, R. E. & Hayes, J. M. *Nature* **359**, 605–609 (1992).

HARDER THAN ROCKET SCIENCE

Dalton Conley is an award-winning researcher who works on the politically charged issues of race, gender and class. He tells **Tony Reichhardt** why he wants to stress the 'science' in the social sciences.

For some in the hard sciences, it was difficult to fathom: a sociologist winning the US National Science Foundation's A. T. Waterman award for achievement by a young researcher? Surely there must be some mistake. But in his acceptance speech last month, Dalton Conley was far from defensive about his discipline's status. "I would like to argue that sociology is among the hardest sciences of all — harder than the proverbial rocket science," he told a black-tie gathering in Washington DC.

Controlled experiments are out, Conley reminded the assembled scientific leaders, and causality is often masked by a confusing tangle of variables. And as if that wasn't difficult enough, sociologists' subject matter includes "the most politically charged and most personally sensitive topics one could address". Some researchers respond to the challenges by retreating into largely descriptive approaches, but Conley has won respect for applying rigorous, quantitative methods. He is the first sociologist to win the prestigious Waterman prize, worth a handsome \$500,000.

Weeks after the award ceremony, in the Washington Square office where he directs New York University's Center for Advanced Social Science Research, Conley says he hasn't decided what to do with the prize money. But relaxing in a room comfortably cluttered with books — including the four he's already written by the tender age of 35 — he is happy to elaborate on the message of his speech. "A lot of the most interesting social science tries to push beyond descriptive accounts of how the world looks and get into causal accounts of how it works," Conley explains. But lacking the ability to manipulate variables the way a



Mix it up: a childhood steeped in contrasts of race and class helped shape Dalton Conley's work.

physicist or chemist would, "you have to look for tricks, natural experiments".

When it comes to finding clever new ways to mine sociological data, Conley admits to being competitive. "I like to squeeze a little bit more out of the lemon after everyone else has squeezed it," he says. One database he's squeezed particularly hard is the Panel Study of Income Dynamics (PSID), a longitudinal study of nearly 8,000 US families. The PSID, administered by the University of Michigan, has been following the same families and individuals since 1968, the year before Conley was born, collecting a wide range of data on their economic status, health, work and family life.

The research problems that interest Conley most have to do with accidents of birth such as skin colour and body size, and how they can affect not just one person, but a whole family tree. His research has advanced our understanding of race and class. And given Conley's own early upbringing on a low-income housing project, he can claim more street credibility than most of his academic colleagues. But what really separates Conley from his peers is his sophistication in analysing multiple variables to tease

out underlying, and sometimes unexpected, patterns.

He has used the PSID to ask who gets ahead and who doesn't in America, and he is especially good at weeding out factors that distract from the real story, such as current annual income. Net worth, he has found, is a much better measure of the resources available to a person, since it captures wealth handed down from earlier generations. As Conley reported in his 1999 book *Being Black, Living in the Red: Race, Wealth, and Social Policy in America*, this accumulated wealth is crucial in determining future economic success.

In black and white

Conley's diligent number-crunching has also provided new insights into the real differences — or lack thereof — between races. While studies have consistently shown that African Americans are more likely to drop out of high school and college, most have ignored the fact that a typical white family has eight times the accumulated wealth of its black counterpart. When Conley applied statistical techniques to remove the variable of family wealth, the most recent cohort of African-Americans in the PSID was actually more likely to graduate from high school than were whites. In an economy where wealth begets wealth, this could be significant

"I like to squeeze a little bit more out of the lemon after everyone else has squeezed it."



K. ALFORD/PANOS PICTURES

out the real costs of unequal opportunity.

Conley's own background provided an education in race, class and the dickensian twists and turns that a life can take. Born to middle-class white parents, he grew up — in part due to his parents' bohemian preferences — in a low-income neighbourhood where nearly everyone else was black or Latino. It was a rough environment, in which a playground quarrel could end with a knife being pulled. When he was 13, a black friend was paralysed by a ricocheting bullet, a traumatic experience that left Conley with an obsessive-compulsive disorder that lasted into adulthood.

Divided lives

Even as a child, however, Conley was aware that he enjoyed privileges that his playmates didn't share. His parents were able to navigate their children through New York City's inequitable educational system — he went to a mostly white, upscale school in Greenwich Village a few blocks from his current office; his own two children attend the same school today. Rather than ending up in a gang like some of his friends, Conley was a semi-finalist in the high-school Westinghouse Science Talent Search. Although he claims never to have made a conscious decision to investigate the factors that shaped his own life, and those of his childhood friends, today he says that it's obvious to him why he later gravitated towards the study of race and class.

As an undergraduate at the University of California, Berkeley, Conley briefly considered a career in biomedical research, before deciding too much of the work he saw was "all trees and no forest". While conceding that sociology can sometimes suffer from the reverse problem, he has no regrets about his chosen field, which allows him to apply quantitative rigour to vexing social questions.

Conley examined one such question — the consequences of birth order — in his 2004 book *The Pecking Order: Which Siblings Succeed and Why*. It's not as simple as the firstborn

always wins, he has found. In fact, using data from the PSID, the General Social Survey and the US Census, Conley concludes that birth order only really counts in families with more than two children — and that external societal and economic pressures play a larger role than expected. Poor people with larger families, he suggests, often are forced to make hard decisions about which children to invest their time and money in — decisions that wealthier parents can avoid.

Lately, Conley has become interested in how body size affects economic fortunes. Using the PSID to include data on older women than other studies have considered, he finds that being overweight leads women to have less workplace prestige, whereas overweight men suffer no such problems. Overweight women also have lower family income, are less likely to get married, and are more likely to be divorced, separated or widowed.

Just add genetics

Conley has always been interested in how biological and social factors interrelate, and he sees big research opportunities in combining genetic and sociological data. Social scientists have been rightly sceptical of the easy explanations for human differences offered by some geneticists, says Conley. "But they're standing on the sidelines carping like little yappy dogs," he argues. "We need to dive headfirst into this." To Conley, genetic differences are just another variable to throw into the mix along with socioeconomic factors.

While many sociologists prefer to think of their field as one of the humanities rather than a science, Conley seems happy either way. Narrative is central to his work, and his books blend quantitative arguments with personal stories. Having written a childhood memoir, *Honky*, at 30, he says he may someday try his hand at a novel, and he alternates writing scientific papers with popular books and op-ed pieces. Sally Hillsman, executive officer of the American Sociological Association, calls Conley "perhaps the most impressive sociology scholar of his generation", while comparing him to Carl Sagan and Stephen Jay Gould in his ability to communicate to a wider audience.

But even with Conley's gift for mixing storytelling and statistics, sociology can never fully account for all of the variables that shape an individual's life. As he wrote in *The Pecking Order*: "Anyone who tells you that they are going to explain your personality, your marriage, your career, or anything else about you with one factor — gender, birth order, income, or astrological sign — might as well be selling you a bottle of snake oil." Conley is no snake-oil salesman, and in that book advocated a more nuanced approach to studying who succeeds, and why: "And just maybe — along the way — we will have a little more sympathy for our less fortunate brothers and sisters." ■

Tony Reichardt is a contributing correspondent for *Nature* in Washington DC.

for social policy. Indeed, Conley's data lead him to suggest that affirmative action should be based on economic class instead of race.

He hasn't come out in favour of paying reparations to the descendants of slaves to right the economic wrongs still being visited upon them — as divisive a topic as one can imagine in American politics — but he doesn't dismiss the argument out of hand. Even if society won't make such payments, he notes, doing the numbers is still worthwhile to point



BUSINESS

Toyota's production line leads from lab to road

Sparks of inspiration and close relations with group companies keep Toyota's central research labs motoring along, as Ichiko Fuyuno reports.

From the outside, Toyota's central research and development laboratories don't look like the intellectual powerhouse behind the world's most successful volume-car producer. Nestled in a collection of unimposing, low-rise buildings between a highway and a maglev railway, the laboratories are thinly populated and dimly lit: managers keep half the lights switched off to save electricity.

But behind the doors of this complex near Nagoya, 370 kilometres west of Tokyo, one of the most highly rated industrial research operations in the world is quietly ticking over.

In contrast with researchers at certain other industrial labs, some of the 700 chemists, physicists, materials scientists and engineers here openly publish their work in journals such as *Nature*. And innovations in materials, electronics and other technologies soon show up in production vehicles.

Analysts say that Toyota has mastered the art of transferring scientific acumen into its cars. "It knows how research is linked to the overall business," says Akinobu Okuda, a specialist in industrial strategy at the Mitsubishi Research Institute in Tokyo.

The laboratories are actually owned jointly by nine Toyota companies, which make everything from cars to forklift trucks. Most of the research is driven by the demands of engineers in these companies.

"We never end with basic research only," says Takashi Saito, a 54-year-old metallurgist and director of the laboratory materials department since 2003. "We always try to envisage final products."

The laboratory follows the Japanese corporate model, providing jobs for life in return for loyalty. Even senior employees are expected to tidy their workspace once a week. Some 20 to 30 researchers are hired

each year. About half are new graduates, the rest come from universities or industry. Managers admit to concerns that the average age of its researchers is increasing. And women account for less than 10% of the research staff, only a dozen or so of whom are non-Japanese.

Researchers talk with group companies almost every day, to keep the ideas rolling. Kazumasa Takatori, head of the inorganic materials laboratory, says that these discussions — rather than journals or conferences — are the best way of staying ahead of rivals.

Smooth ride

The whole operation isn't madly expensive. The central lab spends just over 1% of Toyota's ¥770 billion (US\$7 billion) annual research and development budget. Many research findings are kept in-house: results are only published after consultation with laboratory managers and group companies, and few areas of the laboratory are shown to visitors.

The dedication of the staff is obvious and the results have been impressive, outside analysts say. "The pride that researchers are taking an important role in the Toyota group is perhaps behind their achievements," says Yoshio Watanabe, an auto analyst at Mizuho Securities in Tokyo.

The central laboratory developed, for example, a transistor that controls the electric motor in the Prius hybrid car, which draws power from both an internal combustion engine and an electric motor. It also developed technology to reduce vibration and make the Prius a more comfortable ride.

Laboratory managers play down their contribution to the Prius. But its success is an example of Toyota's conservative yet relentless approach in bringing new technology to market. It began developing the hybrid technology in 1994, when energy prices were low, commercial prospects remote, and few other car makers were taking the idea seriously. Within three years, Toyota had rolled out the world's first hybrid car. At first, the car lost money, but

Forward thinking: first conceived in 1994, the Prius is now the world's leading hybrid car.



Turbo-charged: the achievements of Toyota's Central R&D Laboratory belie its low profile.

growing demand for green vehicles has made it a success in many markets, including the United States, where buyers have to wait months for a vehicle.

Meanwhile, US manufacturers spurned hybrid technology to explore cars based on fuel cells, a more technically ambitious approach that has yet to reach commercial viability. On 14 May, officials from Toyota and General Motors (GM) agreed to collaborate more closely on environmental technologies.

Researchers at the laboratory say they never lose sight of the need to keep costs down. "We should have a business sense," says Saito. In 1998, he recalls, he and his colleagues discovered a titanium-based alloy with higher elastic deformation than any other (T. Saito *et al. Science* **300**, 464–467; 2003). It found a commercial use as a frame for spectacles.

Saito knew the alloy, dubbed 'gum metal', had more potential uses, but it was too expensive — ¥100 million per tonne. He travelled to China ten times, even reaching Mongolia, in his quest for cheap niobium and tantalum to reduce the cost. The new alloy is now also used in golf shafts, and application in cars is under consideration. "It was a hard job," he says.

Requirements are often delivered quite bluntly to staff. Takatori remembers shuddering one summer day in 2000, when the president at a Toyota company that makes car components requested that his group develop a lead-free piezoceramic within two years.

"I thought it'd be impossible to do in 100





TCRDL

years,” says the 51-year-old laboratory manager. Piezoceramics — which convert electricity into mechanical motion, or vice versa — are used in actuators and other car components. But they contain toxic lead. Scientists around the world had struggled to remove the metal without sacrificing performance.

Just-in-time ceramics

But Takatori and his team developed, with some help from the client, a lead-free ceramic that performed just as well as established materials. It was a close call: the project didn't come to fruition until two months before the deadline, when one team member noticed some crystals of an experimental material that were shinier than usual — and delivered ten times more piezoelectric performance than previous tests. Details of the material, a blend of sodium, tantalum and other chemicals, were published as soon as they had been patented (Y. Saito *et al.* *Nature* **432**, 84–87, 2004).

Such innovations, along with the company's manufacturing prowess, have fuelled the ascendancy of Toyota's car division. In the year ended 31 March, Toyota made a record group net profit of ¥1.2 trillion, more than GM, Ford and DaimlerChrysler combined, and a contrast with the loss that GM announced for the first quarter of 2005. Some analysts expect Toyota to displace GM as the world's largest car maker within five years.

Saito says that the laboratory has had a low profile — but that more will be expected of it as Toyota becomes more innovative. “We'll have to be able to produce more creative things from scratch,” he predicts. ■

IN BRIEF

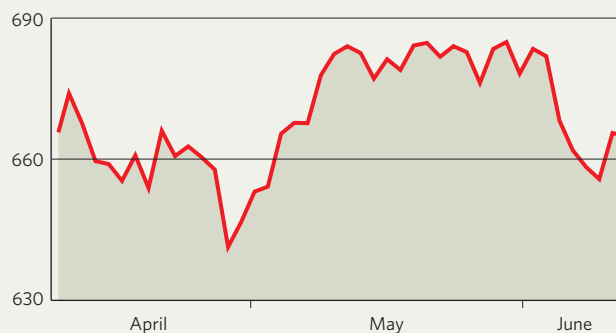
INTEL INSIDE CHINA The world's largest semiconductor manufacturer is establishing a \$200 million venture-capital fund to support emerging companies in China. Intel, which has already invested in dozens of small Chinese companies through an established international venture fund, said the new fund will support companies in hardware, software and computer services. The chip maker says it expects its investment to stimulate innovation in the region, and nurture potential suppliers and customers — as well as making money.

ADVERT AVERSE Television commercials and other advertisements aimed directly at patients are to be dropped by Bristol-Myers Squibb during a product's first year. The New York-based company also pledged to run all of its consumer advertisements past the US Food and Drug Administration for comment, with immediate effect. The announcement reflects growing unease in the drug industry about the cost of mass-market advertising — and its potential to hurt the industry's public image.

PATENT CHANGE FLOATED A proposed reform to US patent law would give precedence to the first person to file for a patent, rather than the first to invent — bringing it in line with most of the rest of the world. The software industry favours the bill, which is sponsored by Lamar Smith (Republican, Texas), chair of the House subcommittee on intellectual property, and the US Patent Office supports parts of it. But other players, including universities and the biotechnology industry, are less enthusiastic. The bill is the opening shot in what could be a multi-year effort to get patent reform through Congress.

MARKET WATCH

Biotechnology stocks



This week Wood Mackenzie, an Edinburgh-based research and consulting firm, reviews recent trends in biotechnology stocks.

After sliding all year, the Nasdaq Biotechnology index steadied in early April, and rose modestly before falling back again last month. The levelling-out reflected good news in the sector, including a couple of large acquisitions and some promising clinical trials results.

In the first acquisition, announced in late April, Shire Pharmaceuticals of Basingstoke, UK, agreed to buy Transkaryotic Therapies for about \$1.6 billion. This company, based in Cambridge, Massachusetts, has both late-stage candidates and approved products for treating rare diseases, which will supplement Shire's portfolio. Then GlaxoSmithKline agreed to buy Corixa, an immunotherapeutics company in Seattle, for \$300 million.

Shares in Transkaryotic and Corixa each rose sharply on the acquisition news.

In mid-May, Cambridge-based Vertex announced positive results in early clinical trials of its drug candidate VX-950 for treating hepatitis C infection. The May meeting of the American Society of Clinical Oncology heard positive news on oncology treatments from several companies including Genentech, whose two drug candidates, Avastin and Herceptin, are showing survival benefits in breast cancer. Although not listed on the Nasdaq, Genentech's exceptional performance helped reinforce investor confidence in biotechnology.

But not all companies fared so well. Shares in OSI Pharmaceuticals of Melville, New York, fell 17% over the two months as investors lost confidence in the approved cancer drug Tarceva, which the company developed with Genentech.

Turkish research council is proud of its independence

SIR — Your recent News story “Turkish government accused of hijacking boosted science budget” (*Nature* **434**, 1055; 2005), is unfair to the managers of the Scientific and Technical Research Council of Turkey, TÜBITAK.

Your sources claim that we are being instrumental in the Turkish government’s supposed “attempts to channel a growing science budget towards [its] supporters”. As acting president of TÜBITAK, I must say that its managers are proud of our 40-year-old council. We know that it has never been and never will be as fragile as some would like to have us fear.

Your News story accurately states: “In many countries, including the United States, governments appoint the officials who run the institutes that distribute public science funding.” Indeed, in some countries, science is funded through a government ministry, and it is only natural that governments, elected by the taxpayers, manage taxpayers’ money. However, what must be absolutely beyond government control is the process of selection and management of funded research projects. This must be done, we agree, through a “robust system of peer review”.

Potential appointments by the government under the proposed law (if it is passed) that are at the centre of the concerns cited in your News story are to be made to the executive board of TÜBITAK: which is just that, a body in charge of administrative, not scientific, decisions. Neither the board nor TÜBITAK’s managers have any say whatsoever in which research projects get funded. The panels of experts in the respective scientific areas carry out evaluation and selection. How those experts are selected is also public knowledge.

Whatever the law, the responsibility of TÜBITAK’s professional managers is to establish an objective and transparent system for evaluating, selecting and monitoring the research projects submitted for support. Indeed, this has been our highest priority since we began work here one-and-a-half years ago. The basic structure of our new system has been open, since October 2004, to public scrutiny and critique at our website, www.tubitak.gov.tr (an English version is under construction).

Moreover, the 200 young scientists mentioned in your News story were selected through this new system, and we are prepared to account for every step of that selection process. We have processed 2,260 projects (a number that has tripled since 2003) by using 778 independent evaluators/referees coming from over 70 universities and institutes — a highly dispersed

distribution of expertise, as planned. This is only the beginning of a healthy trend.

Nüket Yetis

TÜBITAK, Atatürk Bulvarı 221 Kavaklıdere,
06100 Ankara, Turkey

No political agenda in academic bill of rights

SIR — Your News story on our organization’s academic bill of rights, “Professors bristle as states act to mould lecture content” (*Nature* **434**, 686; 2005), quotes the president of a Florida faculty union who claims that the bill would amount to “a right-wing political takeover of the universities”. A representative of the American Association of University Professors (AAUP) is quoted saying it would “politicize the agenda” of higher education.

Yet despite lengthy interviews with three representatives of our organization, your story failed to quote a single staff member, thus denying us the opportunity to respond.

As we repeatedly explained in interviews, the charges raised by faculty opponents and the AAUP are not based on any evidence and represent gross distortions of the bill. Neither the academic bill of rights nor any of the state legislation inspired by it are informed by a political agenda. This is clear from the text itself, which explicitly prohibits the consideration of politics in hiring and tenure decisions and forbids professors to grade students on their political or religious beliefs.

In addition, the bill’s provision requiring professors to make students aware of “serious scholarly viewpoints” other than their own in class echoes the existing policies of the American Historical Association and many public universities. Fringe views, or views based on non-scientific texts such as creationism, could not be considered “serious scholarly viewpoints”.

We invite readers of *Nature* to visit our website, www.studentsforacademicfreedom.org, and to form their own judgments.

Sara Dogan

Students for Academic Freedom, 1411 K Street, NW,
Suite 1100, Washington, DC 20005, USA

Sale of public databases puts biological data at risk

SIR — For many years, biology journals have expected data to be submitted to public databases before publication. For example, requiring a GDB accession number (or D-segment identifier) before publication of human gene mapping results enhanced the growth of GenBank. When these policies were put in place by journal publishers, they were based on the reasonable assumption that biological databases would be constantly

supported and updated, and that the data would always be freely available.

With the number of biological databases growing in proportion to the growth of data, it is reasonable to expect that natural selection will occur: the best resources will thrive and others will become extinct. It is during the extinction process that accessibility to submitted data will be in question.

On 18 April 2005 a press release from Blueprint Initiative (which is the umbrella organization covering the BIND database; www.bind.ca) announced that Nature Publishing Group would “submit manuscripts containing biomolecular interaction data to the BIND database in advance of publication” and “in a manner similar to the publication of GenBank identifiers for publications containing novel sequences.” But then on 2 May 2005 Blueprint announced that it would start winding down its North American operations.

To its credit, the Blueprint group claims that the BIND database will continue to be accessible and curated from its Singapore offices. However, the long-term future of BIND remains in question, and it is up to journal publishers to determine their future relationships with it.

“A scenario where data could be lost, or even sold off commercially, is not impossible to imagine.”

— A. Jamie Cuticchia, Gregg W. Silk

We welcome assurances from the Blueprint Initiative that BIND data will continue to be made freely available. However, the funding situation in which BIND has been placed is a warning of what could conceivably happen to databases. A scenario where data could be lost, or even sold off commercially, is not impossible to imagine.

Whether scientists will continue to ardently support journals selectively submitting pre-publication papers to databases, or requiring authors to submit data themselves prior to publication, will remain to be seen. Whatever happens to BIND, assuring the availability and rights to the scientists’ own data should remain a number-one priority of both journals and database managers.

A. Jamie Cuticchia, Gregg W. Silk

Bioinformatics, Research Triangle Institute,
3040 Cornwallis Road, Research Triangle Park,
North Carolina 27709, USA

No NPG journal insists that authors submit data to BIND in advance of publication. Some NPG journals send manuscripts to BIND shortly before publication so that accession number links can be included in the paper, but authors can choose to opt out of this system. For all other journals (including *Nature*) BIND curates the data after publication — Editor, *Nature*.

BOOKS & ARTS

Switching on evolution

How does evo-devo explain the huge diversity of life on Earth?

Endless Forms Most Beautiful: The New Science of Evo Devo

by Sean B. Carroll

W. W. Norton: 2005. 350 pp. \$25.95

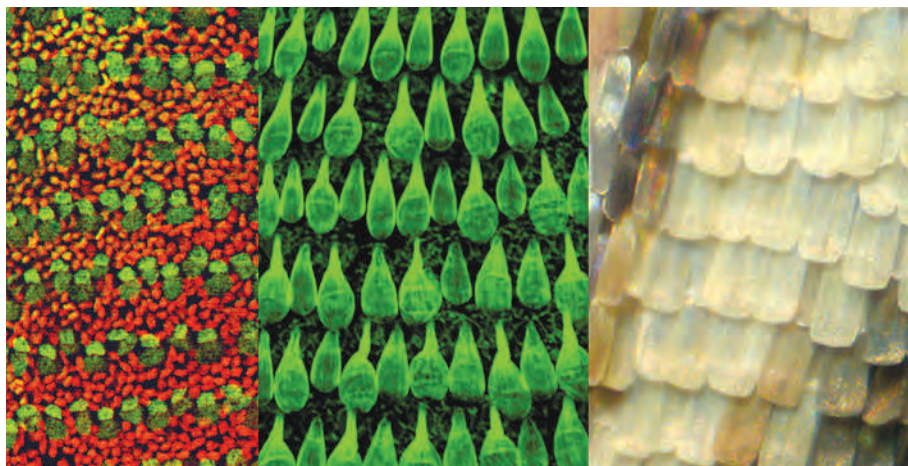
Jerry A. Coyne

Aimed at the interested lay reader, *Endless Forms Most Beautiful* (the title comes from the last paragraph of Darwin's *Origin of Species*) is a paean to recent advances in developmental genetics, and what they may tell us about the evolutionary process. The book's centrepiece is the unexpected discovery that the genes that control the body plans of all bilateral animals, including worms, insects, frogs and humans, are largely identical. These are the 'homeobox' (*Hox*) genes, whose products bind to the DNA of other genes, triggering a cascade of processes that ultimately yield eyes, limbs, hearts and other complex structures.

The evolutionary conservatism of these genes across long-diverged species is staggering. Only a jaded biologist could not be astonished at the ability of the *Pax-6 Hox* gene from mice (which triggers eye formation) to induce in the fruitfly *Drosophila* the formation of fly eyes all over the body, even on the wings. Remarkably, *Pax-6* helps to organize compound eyes in flies and camera eyes in both squid and vertebrates — structures once thought to have evolved independently. Another *Hox* gene, *tinman*, induces heart formation in both insects and vertebrates, and *Distal-less* controls the development of fly legs, fish fins and the tube feet of sea urchins.

Sean Carroll, a leader in the field of evolutionary developmental biology (evo-devo), is an adept communicator, conveying the intricacies of development in clear and lively prose. He ranges across the history of biology, linking the early concept of 'inducers' to today's more complex view of developmental networks, and explores the implications of evo-devo for the Cambrian explosion, the biology of dinosaurs, the brains of humans, and the striping of zebras.

Endless Forms Most Beautiful is a first-rate introduction to evo-devo for the scientifically curious, but its faintly self-congratulatory message — that the most important problems in understanding the evolution of development have been solved — left me feeling uncomfortable. Carroll presents his vision of the field without admitting that large parts of



Differential gene expression, cellular protrusions and scales combine to make butterfly wings.

that vision remain controversial. I would have appreciated a caveat or two, and non-scientists may mistakenly believe that Carroll presents the scientific consensus about evolution and development.

Carroll emphasizes throughout that the evolution of animal form and complexity results from three factors. The first is modularity of organization: the ground plan of bilateral animals involves repeated segments that can evolve independently. The lobster, for example, is a veritable 'Swiss Army' crustacean, whose diverse appendages — antennae, mouthparts, claws, walking legs, swimming legs and tail — are all modified ancestral limbs. The second factor is that most animals share a small but similar set of 'tool-kit genes' that regulate the development of different modules. These genes, which produce regulatory proteins called transcription factors, are highly conserved in function; *Hox* genes are the canonical example.

But modularity and a shared genetic tool kit cannot by themselves account for "endless forms", because conserved genes cannot explain diversity. Carroll therefore repeatedly emphasizes his third thesis: that the main engine of evolution is not change in protein-coding genes but in the switches that control them. Changes in these switches — the promoters and enhancers in DNA that regulate the transcription of protein-coding genes — supposedly promote evolution by causing existing genes to be expressed at new times and places. This idea has been with us for a

long time. Around 1970, the biologists Roy Britten, Eric Davidson and Allan Wilson were already arguing that the 'regulatory gene' is the locus of evolution, and the idea is now accepted wisdom among evo-devotees.

The evidence for this critical hypothesis, however, rests more on inference than on observation or experiment. Carroll first notes that dissimilar species can in fact be genetically similar: "Mice and humans have nearly identical sets of about 25,000 genes" and "chimps and humans are almost 99 percent identical at the DNA level. Since the sets of genes are so widely shared, how do differences arise?" His answer is the evolution of non-coding regulatory elements: whether you are a man or a mouse apparently depends solely on your promoters and enhancers. But the underlying statistics are deceptive; even a 1% difference in DNA sequence implies a substantial difference in protein sequence. We now know that humans and chimps have different amino-acid sequences in at least 55% of their proteins, a figure that rises to 95% for humans and mice. Thus we can't exclude protein-sequence evolution as an important reason why we lack whiskers and tails.

Carroll also claims that proteins are resistant to evolutionary change: they are often involved in many pathways, and therefore a change in protein sequence, while enhancing one aspect of the protein's many functions, could damage several others. In contrast, changing an enhancer or promoter can affect the expression of a single protein without altering its

S. PADDOCK, J. GATES, S. CARROLL/HIMI/UNIV. WISCONSIN

structure, so such changes are more likely to be adaptive. He deduces that “the evolution of ‘new genes’ is not the explanation for the origin of diversity of most animal groups”. Rather, “it is the switches that encode instructions unique to individual species and that enable different animals to be made using essentially the same tool kit,” he says. “Evolution of form is very much a matter of teaching very old genes new tricks!”

But recent data cast doubt on this argument. Humans have about 32,000 protein-coding genes, fruitflies only 13,000. Clearly, the difference between these species involves the origin of new proteins: in fact, between 40% and 50% of our protein-coding genes have no known homologues in flies. So one could argue that the evolution of form is very much a matter of teaching old genes to make new genes. And, given the data, this cannot be difficult.

There are several ways that protein structure can evolve without injurious side effects. One of the most common is gene duplication. Extra copies of a gene can arise by unequal crossing over or by reverse transcription, allowing one copy to retain its function while the other assumes a new function. This process has been a major force in evolution. A large fraction of genes (at least 39% in humans) are members of families derived from repeated duplications and diversification of ancestral genes, a process that has yielded many evolutionary novelties. These families include the globins (such as myoglobin and the various haemoglobins); immunoglobulins; opsins (which led to colour vision in Old World primates); and olfactory receptors (almost certainly involved in the evolution of a keen sense of smell in land animals). Lactalbumin, which helps to produce milk in mammals, resulted from a duplication of lysozyme, and the crystallins of our eye lenses are ultimately derived from heat-shock genes.

This ‘multiply and diversify’ model of molecular evolution does not depend solely on the duplication of individual genes; the evolution of tetrapods apparently involved at least two bouts of whole-genome duplication. Many evolutionists agree with the geneticist Wen-Hsiung Li’s conclusion that “there is now ample evidence that gene duplication is the most important mechanism for generating new genes and new biochemical processes that have facilitated the evolution of complex organisms from primitive ones”. Carroll, however, seems too enamoured of his ‘regulation is all’ thesis to consider this alternative view.

There are other ways beside gene duplication that proteins have evolved adaptively. These include gene conversion, recruitment of genes to new functions (responsible for creating the antifreeze glycoproteins that allow fish to live in frigid waters), exon shuffling (involved in the evolution of blood clotting factors) and the addition of transposable elements to coding sequences. Finally, and

simplest of all, we have many examples of adaptive changes of protein sequence between closely related species, including differences in the coat colour of mice, the digestive enzymes of herbivores, and the haemoglobins of high-altitude birds and mammals.

In contrast, the evidence for the adaptive divergence of gene switches is still thin. The best case involves the loss of protective armour and spines in sticklebacks, both due to changes in regulatory elements. But these examples represent the loss of traits, rather than the origin of evolutionary novelties. Carroll also gives many cases of different expression patterns of *Hox* genes associated with the acquisition of new structures (such as limbs, insect wings and butterfly eyespots), but these observations are only correlations. One could even argue that they are trivial. Given the centrality of *Hox* genes in development, it is almost inevitable that such genes are involved in the evolution of a new trait. Carroll’s correlations, however, do not compel us to believe that changes in these genes are the key factor in the evolution of such traits. We now know that *Hox* genes and other transcription factors have many roles besides inducing body pattern, and their

overall function in development — let alone in evolution — remains murky.

In the end, we simply don’t know the relative importance of protein and non-protein changes in creating biological diversity. In many cases, both must have evolved in tandem, as different members of gene families are often expressed in different tissues or at different times. For example, the protein sequence of fetal γ -haemoglobin evolved adaptively to wrest oxygen from the mother’s blood, but its gene is turned off after birth, probably by new regulators. Carroll’s emphasis on gene switches may prove correct, but this awaits the labours of the next generation of biologists.

Although *Endless Forms Most Beautiful* is a lucid and valuable summary of evo-devo, it does proclaim a clever but still unproved hypothesis as central to the evolutionary process. As Carroll himself notes: “Simplification may indeed be necessary for news articles, but it can distort the more complex and subtle realities of evolutionary patterns and mechanisms” ■

Jerry A. Coyne is in the Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA.

EXHIBITION

Fresh flowers

A New Flowering: 1000 Years of Botanical Art

At the Ashmolean Museum in Oxford, UK, until 11 September 2005.
www.ashmol.ox.ac.uk/ash/exhibitions/exh075.html

Colin Martin

Botanical artists face the dual challenge of capturing the essence of each plant species artistically, yet representing the plants and their stages of growth with absolute scientific

accuracy. The veracity of great botanical art derives from artists’ skill at producing work that both records plant species in technical detail and heightens viewers’ perception of the natural world.

The exhibition *A New Flowering* is a clever juxtaposition of examples of contemporary botanical art, selected from a private collection assembled since 1990, with rarely exhibited works from the past millennium, chosen from the collections of the Ashmolean Museum and several Oxford colleges. It highlights both

the novelty of current practice in botanical art and its continuity with the past.

Historic works on show include eleventh-century herbals, which name and describe plants and list their properties and uses; fifteenth-century illuminated manuscripts, whose borders are decorated with flowers; sixteenth-century printed books, illustrated with woodcuts; seventeenth-century books that are illustrated with engravings; and eighteenth- and nineteenth-century works, notably some particularly sumptuous volumes that depict freshly discovered plant species. Throughout, contemporary botanical images are displayed alongside related earlier works.



Setting the scene: the Scottish island of Ailsa Craig is the backdrop to Rory McEwen’s painting of monk’s head.

Two important twentieth-century botanical artists whose works are included in the exhibition expanded their influence beyond the world's herbaria into environmental conservation and modern art, respectively. English artist Margaret Mee moved to Brazil in the 1950s, where she made 15 expeditions into the

Amazon rainforest and became one of the first conservationists to raise concerns over its exploitation. And Scottish artist Rory McEwen, for whom the placement of flowers, vegetables and leaves in his pictures was as important as the plants themselves, propelled botanical art into modern art galleries, includ-

ing the Museum of Modern Art in New York. His eerie 1977 watercolour of an unfolding young shoot of monk's hood (*Aconitum*) with a ghostly profile of the Scottish island Ailsa Craig in the background, shown on the opposite page, is particularly striking. ■

Colin Martin is a writer based in London.

Flood warnings

The Future of Large Dams: Dealing with the Social, Environmental and Political Costs

by Thayer Scudder

Earthscan: 2005. 432 pp. £45

Christer Nilsson

During the past 50 years the world has experienced an unprecedented increase in the number of large dams, from 5,700 in 1950 to approximately 50,000 today. The International Commission on Large Dams defines a large dam as one spanning 15 metres or more from base to top, or storing more than 3 million cubic metres of water. Almost half of them are in China, but the United States, India, Spain and Japan also have large numbers. The peak years for dam building were the 1970s, but construction continues apace, especially in countries that were late to industrialize.

But how many large dams do we really want? Should old ones be decommissioned and plans for new ones halted because they cause serious, irreversible degradation of life-support systems? For example, dams can eliminate ecosystems and reduce biodiversity, fisheries and bioproduction. Or should we keep our dams because they provide necessary water and energy to populations that have grown beyond the carrying capacity of their environments? These are difficult political questions, and Thayer Scudder adds a human dimension that is surprisingly often ignored or inadequately understood in the planning and building of large dams. In *The Future of Large Dams* he speaks up for the millions of people who are directly affected by large dams and reservoirs, many of whom are "poor, uneducated, relatively powerless rural residents".

Scudder, an anthropologist at the California Institute of Technology, writes with great authority, having been one of the 12 commissioners of the World Commission on Dams. This role has given him a remarkable insight into the global politics of large dams and their social, economic and environmental consequences. His position on large dams has changed from strong support to anger and despair over the poor treatment of people who live near dams. He now believes that most of the really large dams impose unacceptable environmental and social costs.

Hundreds of millions of people are adversely affected by dams, but curiously no precise figures are available. At least 40–80



Water power: large dams such as China's Shuikou Dam can have a huge effect on local populations.

million people have been resettled from planned reservoir basins, and this evacuation also affects the populations who receive them. Additionally, the lifestyles of those living downstream of the dam are changed because the water flow is regulated. Scudder spends a large part of his book discussing the social complexities of these three groups of people.

He discusses the plight of people who have lived for generations close to a free-flowing river but who may suffer from shortage of food and water, or contact with new diseases, after relocation. People may be stressed by having to leave burial grounds or religiously important sites, and local leaders may be undermined whether they oppose resettlement or not. Scudder presents a new study on people whose parents were resettled, involving a total of nearly 1.5 million people affected by 50 different dams across the world. Their living standards improved in only 7% of cases but worsened in 70%, with the rest having no significant change. Scudder also provides an in-depth analysis of a few major dam projects and discusses the institutions involved in the politics and economics of large dams. The book's personal tone makes it enjoyable to read, but this is also a scholarly book

that has numerous footnotes and references.

The story that Scudder tells is rather depressing, but he does try to find a way forwards. In the final chapter, he lists seven sets of recommendations on how to make decisions about the possible construction of large dams in the future. These recommendations are derived from suggestions made by the World Commission on Dams and are expanded to take account of existing dams. They include more thorough investigations into the possible consequences of dams, improved stakeholder involvement, and an active search for alternative solutions to water-related problems.

Scudder concludes by stating that the application of his recommendations would considerably reduce the number of large dams being built. Sustainability should be a top priority, and the book is a valuable reminder of the dangers of destroying sustainable rural societies largely to support unsustainable cities or large industries. Hopefully, Scudder's book will help to lessen the damage caused by the building and maintenance of large dams. ■

Christer Nilsson is in the Landscape Ecology Group, Department of Ecology and Environmental Science, Umeå University, Umeå 90187, Sweden.

AP/EMPICS

Dynamic Universe

The first person to carry out a modern survey of the night sky, Fritz Zwicky's astronomical observations led to a new picture of a turbulent Universe that is punctuated by violent events.

Freeman Dyson

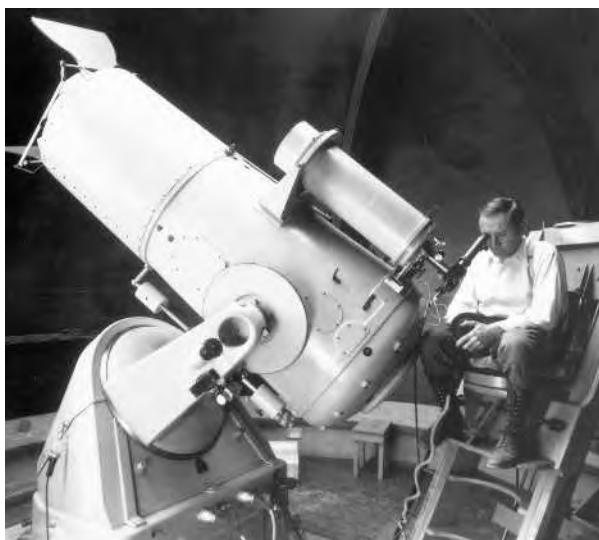
The Swiss physicist Fritz Zwicky (1898–1973) was responsible, more than anyone else, for a profound change in our view of the astronomical Universe. Before Zwicky, the ancient aristotelian view of the celestial sphere as a region of eternal harmony and tranquillity was still largely intact, and the job of an astronomer was to make accurate maps of an unchanging landscape. After Zwicky, the modern view of the Universe emerged, as a dynamic scene dominated by violent events. The job of an astronomer today is to record and interpret the processes of change.

Zwicky's great work was done in the 1930s when he was an associate professor of physics at Caltech, the California Institute of Technology. He had been trained at the Swiss Federal Institute of Technology as an X-ray crystallographer and had no official credentials as an astronomer. But he saw an opportunity arising from two unconnected events that happened almost simultaneously.

First, in 1928, the Rockefeller Foundation awarded a large sum of money to Caltech for the construction of a major astronomical observatory. Second, in 1930, the one-armed optical genius Bernhard Schmidt in Hamburg invented a telescope with a revolutionary new design, which allowed high-resolution photography over a wide field of view. Zwicky's astronomer friend Walter Baade happened to be a personal friend of Bernhard Schmidt and saw the original Schmidt telescope in action in Hamburg. Zwicky heard what the telescope could do and grabbed the opportunity. He persuaded Caltech to buy an 18-inch Schmidt camera and install it at the new observatory on Palomar Mountain. He made sure that he would have full-time use of the camera — at that time the only wide-field camera in the world at a site with good astronomical seeing.

Zwicky understood that to see rare, violent and short-lived events in the Universe, he had to photograph large areas of sky repeatedly. With his little Schmidt camera, he had a unique opportunity to photograph the entire northern sky over and over again. With a single assistant to help him, he continued for four years to survey

the northern sky. Two major discoveries that emerged from his survey were supernovae and dark matter. Zwicky observed 20 supernovae, a large enough sample to allow him to classify them into several types and infer their different modes of origin. His discovery of dark matter came from studying the motions of individual galaxies in rich clusters of galaxies, and from calculating that the visible mass in the clusters was insufficient by a large factor to cause the observed motions.



Using the Schmidt telescope, Fritz Zwicky discovered dark matter.

In his summary of the survey, Zwicky proudly asserted: “for the construction of the 18-inch Schmidt telescope, its housing, a full-size objective prism, a small remuneration for my assistant, and the operational costs for the whole project during ten years, only about \$50,000 dollars were expended. This probably represents the highest efficiency, as measured in results achieved per dollar invested, of any telescope presently in use, and perhaps of any ever built, with the exception of Galilei's little refractor.”

Zwicky's sky survey set the pattern for many later surveys done with bigger instruments and greater investments of manpower and money. The newest sky survey Pan-STARRS, due to begin in 2006, will follow Zwicky's example in emphasizing rapid and repeated coverage of the sky. It should discover a wealth of short-lived phenomena at all distances, from near-Earth asteroids to optical afterglows of gamma-ray bursts in remote galaxies.

Zwicky's radical ideas and pugnacious personality brought him into frequent con-

flict with his colleagues at Caltech. They considered him crazy and he considered them stupid. He never became an accepted member of the astronomical community, but followed his own path. During the Second World War he was director of research at the Aerojet Corporation, which developed rockets for the military. At the same time he organized the Committee for Aid to War-stricken Scientific Libraries. This collected massive quantities of scientific books and journals and distributed them to libraries that had been disrupted or destroyed during the war. He founded the committee in 1941 and ran it with his habitual enthusiasm and efficiency. For this work he received the Medal of Freedom from President Harry Truman in 1949.

Besides his revolutionary work as an observer and organizer, Zwicky made revolutionary contributions to theoretical astronomy. He published papers about neutron stars, supernovae, black holes and gravitational lenses, long before these subjects became fashionable. His thinking was based on a personal philosophy which he called the ‘morphological approach’. The idea is that you first make a complete list of all

possible solutions to a problem, and then choose the least unlikely solution for further investigation. By following this approach, you have a good chance of finding new solutions that other people overlooked.

Zwicky applied the morphological approach both to theoretical and practical problems; it was this approach that led him to become the first modern astronomer, with a new and dynamic view of the Universe. In his book *Discovery, Invention, Research, Through the Morphological Approach*, he describes successful application of the morphological approach to a multitude of problems arising in pure mathematics, automobile design and university administration, as well as in rocket propulsion and astronomy. Unfortunately, the approach does not seem to work so well if your name is not Fritz Zwicky. ■

Freeman Dyson is at the Institute for Advanced Study, Einstein Drive, Princeton, New Jersey 08540, USA.

CALIFORNIA INSTITUTE OF TECHNOLOGY

LOW-TEMPERATURE PHYSICS

A quantum revolution

Rudolf Grimm

Tiny quantum tornadoes observed in ultracold gases of fermionic atoms provide definitive evidence of superfluidity, and open up new vistas in the modelling of quantum many-body systems.

Almost exactly ten years after the first observation of a Bose–Einstein condensate (BEC) in ultracold atomic gases consisting of so-called bosons^{1,2}, a similar revolution is now unfolding. Evidence has piled up that atoms of the class of particles known as fermions can also be cooled down to a superfluid state. On page 1047 of this issue, Zwierlein *et al.*³ present a final, spectacular proof for superfluidity — frictionless flow — in an ultracold gas of fermionic atoms.

Fundamental particles are divided into bosons and fermions depending on their internal angular momentum, or ‘spin’. If the total spin is an integer multiple of Planck’s constant, h , divided by 2π , the particle is a boson. An ultracold ensemble of these particles can condense into the lowest possible quantum energy state, where it forms a BEC. The building blocks of matter such as electrons, protons and neutrons are, however, particles with half-integer spin — fermions. Fermions obey Pauli’s exclusion principle, which forbids two or more particles to occupy the same quantum state. The formation of an ultracold condensate similar to a BEC is thus not allowed for a system of single fermions.

Fermions can, however, condense into a macroscopic quantum state and form a superfluid if they pair up, forming compound objects with whole-integer spin and bosonic character. Bardeen–Cooper–Schrieffer (BCS) theory⁴, for example, describes the frictionless

transport of electrons in superconductors in terms of composites known as Cooper pairs. The great interest in ultracold Fermi gases^{5,6} is due to their unique properties for modelling the physics of quantum matter in general — table-top experiments promise insights not only into the mechanisms of high-temperature superconductivity, but also into the physics underlying neutron stars and the quark–gluon plasma, the state of matter thought to have dominated at a critical stage in the early development of the Universe.

A crucial parameter in these situations is the interaction strength between two fermions, as this determines the binding energy and size of the pairs, and thus the macroscopic properties of the quantum system. In an ultracold gas, the pair interaction can be varied conveniently with a magnetic field if a so-called Feshbach resonance is present. Below this resonance, a regime of strong pairing can be realized, in which fermionic atoms bind together to form bosonic molecules that eventually condense into a molecular BEC. Far above the Feshbach resonance, weak Cooper pairing leads to a BCS-type system. The ability continuously to vary the system properties in between these two extremes has opened up the possibility of studying the long-elusive crossover from a BEC to a BCS state. The fundamental properties of this crossover have been studied with lithium (⁶Li) and potassium (⁴⁰K) gases. But although the observation of pair condensa-

tion^{7,8}, and measurements of collective oscillation modes^{9,10}, pairing energy¹¹ and heat capacity¹², together with supporting theory, provided compelling evidence for superfluidity, a final proof — a ‘smoking gun’ — was still missing.

So what is an unambiguous signature for frictionless flow in a macroscopic quantum state? A striking possibility results from the discrete nature of angular momentum in quantum mechanics. A superfluid cannot rotate like a classical fluid, but arranges itself in a system of vortices, with each of these tiny quantum tornadoes carrying a separate chunk of the total angular momentum of the system. The vortices expel particles from their centres, forming filament-like empty cores that penetrate the superfluid. The vortices also repel each other, leading, in thermal equilibrium, to their crystallization into a regular lattice, known as an Abrikosov lattice. Such structures are a well established signature of superfluid flow in some kinds of superconductors⁴, and have also been observed for rotating BECs.

Zwierlein and colleagues’ experiment³ with a rotating Fermi gas posed great challenges. After first using a BEC consisting of sodium atoms to optimize their setup for vortex creation (Fig. 1a), the authors primed a laser trap, carefully optimized to be as round as possible, with an ultracold Fermi gas of ⁶Li atoms. Two additional laser beams, swirling like spoons in the quantum fluid, stirred the lithium gas vigorously to introduce angular momentum

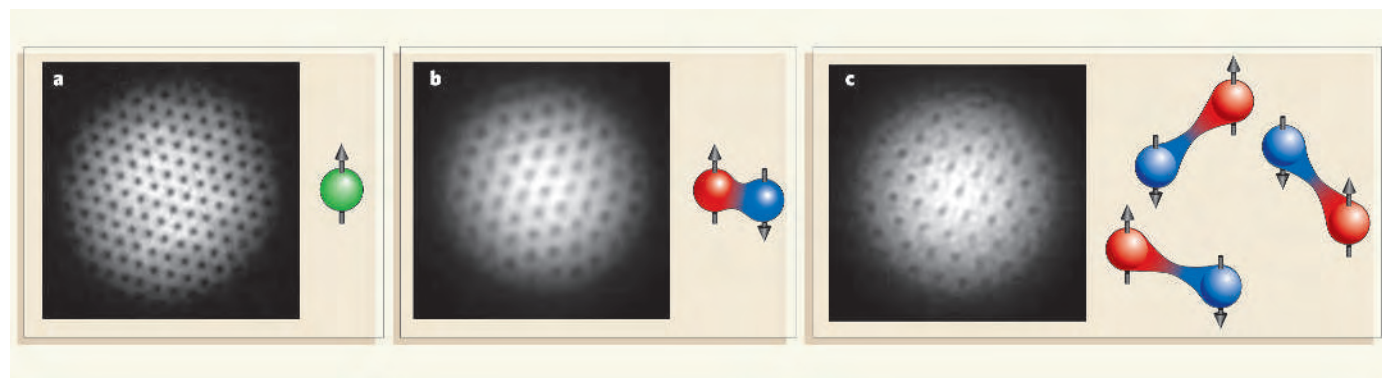


Figure 1 | Cool rotations. Vortex structures observed in rotating superfluids by Zwierlein *et al.*³. **a**, A Bose–Einstein condensate (BEC) of bosonic (integer spin) sodium atoms. **b**, A BEC formed of two fermionic (half-integer spin) ⁶Li atoms bound together tightly to form a gas of bosonic molecules. **c**, A Fermi gas of loosely bound pairs of ⁶Li atoms in the strongly interacting regime, the first unambiguous sign of superfluidity seen in a fermionic gas.

ANDRE SCHIOTZEK

to the system, which was then given a variable time for formation and crystallization of the vortices.

The vortex cores are far too small to be resolved by optical imaging. So Zwierlein and colleagues magnified the vortex cores and the whole vortex lattice by turning off the laser trap and releasing the system into free space, where it expanded. They also increased the size of the vortex cores, and thus their visibility, by changing the interaction strength during the expansion.

The authors first demonstrated the formation of vortex lattices in the lithium gas in the molecular BEC regime. Here the size of the fermion pairs is small compared with the typical interparticle distances, and a closely bound, bosonic molecule is formed (Fig. 1b). In the strongly interacting regime close to the Feshbach resonance on the BCS side, the pair size is comparable to typical interparticle distances. Here, the fermion pairs cannot bind together to form isolated molecules — yet similar vortex patterns were observed (Fig. 1c). The time required for the formation of the vortex lattice was about a hundred times longer than the expansion timescale — ruling out the possibility that vortices are formed during expansion.

The spectacular observation of vortices in a Fermi gas heralds the advent of a new era of research reaching far beyond Bose–Einstein condensation. As an immediate experimental step, interfering light fields can be used to simulate a crystal lattice¹³, providing a unique tool for solving problems in condensed-matter physics¹⁴. And the amazing level of control demonstrated in the work of Zwierlein *et al.*³ can be extended to more sophisticated systems — mixed Fermi systems could be used to simulate a nucleus of protons and neutrons, or exotic superconductors. This final proof of superfluidity in a Fermi system opens fantastic new prospects for many different fields of many-body quantum physics. ■

Rudolf Grimm is at the Institute of Experimental Physics, University of Innsbruck, and the Institute of Quantum Optics and Quantum Information, Austrian Academy of Sciences, 6020 Innsbruck, Austria.
e-mail: rudolf.grimm@ultracold.at

1. Cornell, E. & Wieman, C. *Rev. Mod. Phys.* **74**, 875–893 (2002).
2. Ketterle, W. *Rev. Mod. Phys.* **74**, 1131–1151 (2002).
3. Zwierlein, M. W., Abo-Shaeer, J. R., Schirotzek, A., Schunck, C. H. & Ketterle, W. *Nature* **435**, 1047–1051 (2005).
4. Tinkham, M. *Introduction to Superconductivity* 2nd edn (Dover, Mineola, NY, 2004).
5. Cho, A. *Science* **301**, 750–752 (2003).
6. Chevy, F. & Salomon, C. *Phys. World* **18** (3), 43–47 (2005).
7. Regal, C., Greiner, M. & Jin, D. S. *Phys. Rev. Lett.* **92**, 040403 (2004).
8. Zwierlein, M. W. *et al. Phys. Rev. Lett.* **92**, 120403 (2004).
9. Kinast, J., Hemmer, S. L., Gehm, M. E., Turlapov, A. & Thomas, J. E. *Phys. Rev. Lett.* **92**, 150402 (2004).
10. Bartenstein, M. *et al. Phys. Rev. Lett.* **92**, 203201 (2004).
11. Chin, C. *et al. Science* **305**, 1128–1130 (2004).
12. Kinast, J. *et al. Science* **307**, 1296–1299 (2005).
13. Köhl, M., Moritz, H., Stöferle, T., Günter, K. & Esslinger, T. *Phys. Rev. Lett.* **94**, 080403 (2005).
14. Jaksch, D. & Zoller, P. *Ann. Phys.* **315**, 52–79 (2005).

NEUROSCIENCE

Friends and grandmothers

Charles E. Connor

How do neurons in the brain represent movie stars, famous buildings and other familiar objects? Rare recordings from single neurons in the human brain provide a fresh perspective on the question.

‘Grandmother cell’ is a term coined by J. Y. Lettvin to parody the simplistic notion that the brain has a separate neuron to detect and represent every object (including one’s grandmother)¹. The phrase has become a shorthand for invoking all of the overwhelming practical arguments against a one-to-one object coding scheme². No one wants to be accused of believing in grandmother cells. But on page 1102 of this issue, Quiroga *et al.*³ describe a neuron in the human brain that looks for all the world like a ‘Jennifer Aniston’ cell. Ms Aniston could well become a grandmother herself someday. Are vision scientists now forced to drop their dismissive tone when discussing the neural representation of matriarchs?

A more technical term for the grandmother issue is ‘sparseness’ (Fig. 1). At earlier stages in the brain’s object-representation pathway, the neural code for a given stimulus is a broad activity pattern distributed across a population of neurons, each responsive to some discrete visual feature⁴. At later processing stages, neurons become increasingly selective for combinations of features⁵, and the code becomes increasingly sparse — that is, fewer neurons are activated by a given stimulus, although the code is still population-based⁶. Sparseness has its advantages, especially for memory, because compact coding maximizes total storage capacity, and some evidence suggests that ‘sparsification’ is a defining goal of visual infor-

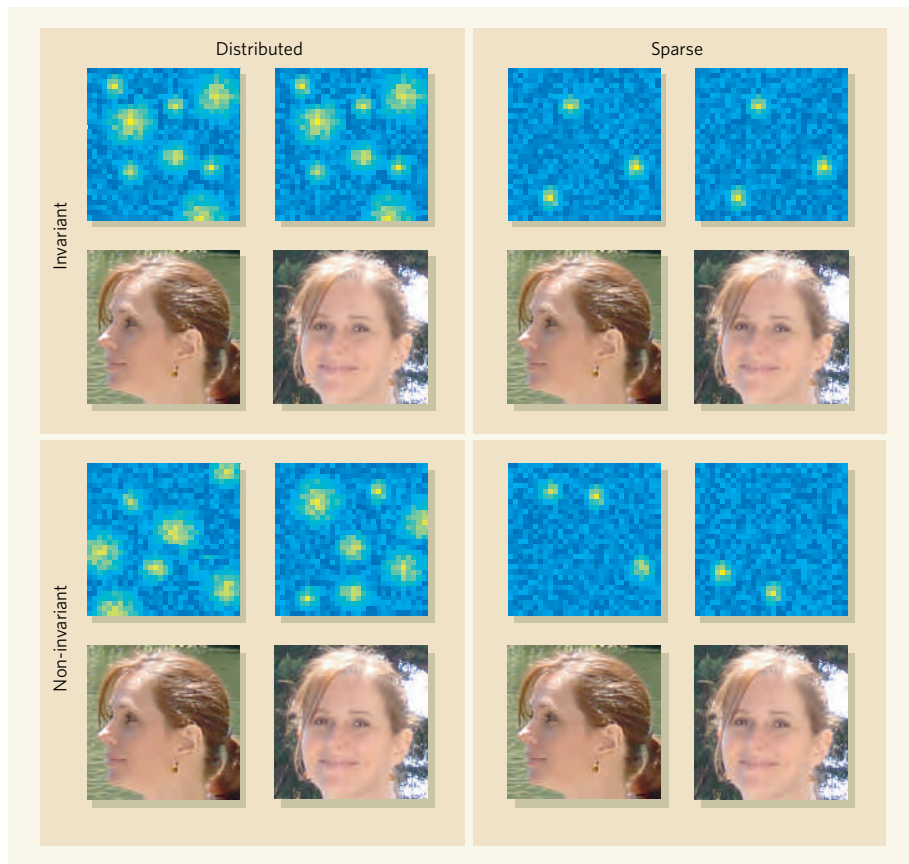


Figure 1 | Sparseness and invariance in neural coding of visual stimuli. The blue and yellow pixel plots represent a hypothetical neural population. Each pixel represents a neuron with low (blue) or high (yellow) activity. In distributed coding schemes (left column), many neurons are active in response to each stimulus. In sparse coding schemes (right column), few neurons are active. If the neural representation is invariant (top row), different views of the same person or object evoke identical activity patterns. If the neural representation is not invariant (bottom row), different views evoke different activity patterns. The implication of Quiroga and colleagues’ results³, at least as far as vision is concerned, is that neural representation is extremely sparse and invariant.

mation processing^{7,8}. Grandmother cells are the theoretical limit of sparseness, where the representation of an object is reduced to a single neuron.

Quiroga and colleagues³ report what seems to be the closest approach yet to that limit. They recorded neural activity from structures in the human medial temporal lobe that are associated with late-stage visual processing and long-term memory. The structures concerned were the entorhinal cortex, the parahippocampal gyrus, the amygdala and the hippocampus, and the recordings were made in the course of clinical procedures to treat epilepsy.

The first example cell responded significantly to seven different images of Jennifer Aniston but not to 80 other stimuli, including pictures of Julia Roberts and even pictures of Jennifer Aniston with Brad Pitt. The second example cell preferred Halle Berry in the same way. Altogether, 44 units (out of 137 with significant visual responses) were selective in this way for a single object out of those tested.

The striking aspect of these results is the consistency of responses across different images of the same person or object. This relates to another major issue in visual coding, 'invariance' (Fig. 1). One of the most difficult aspects of vision is that any given object must be recognizable from the front or side, in light or shadow, and so on. Somehow, given those very different retinal images, the brain consistently invokes the same set of memory associations that give the object meaning. According to 'view-invariant' theories, this is achieved in the visual cortex by some kind of neural calculation that transforms the visual structure in different images into a common format^{9–11}. According to 'view-dependent' theories, it is achieved by learning temporal associations between different views and storing those associations in the memory^{12–14}.

Quiroga and colleagues' results³ set a new benchmark for both sparseness and invariance, at least from a visual perspective. Most of the invariant structural characteristics in images of Jennifer Aniston (such as relative positions of eyes, nose and mouth) would be present in images of Julia Roberts as well. Thus, any distributed visual coding scheme would predict substantial overlap in the neural groups representing Aniston and Roberts; cells responding to one and not the other would be rare. The clean, visually invariant selectivity of the neurons described by Quiroga *et al.* implies a sparseness bordering on grandmotherliness.

However, as the authors discuss, these results may be best understood in a somewhat non-visual context. The brain structures that they studied stand at the far end of the object-representation pathway or beyond, and their responses may be more memory-related than strictly visual. In fact, several example cells responded not only to pictures but also to the printed name of a particular person or object.

Clearly, this is a kind of invariance based on learned associations, not geometric transformation of visual structure, and these cells encode memory-based concepts rather than visual appearance.

How do you measure sparseness in conceptual space? It's a difficult proposition, requiring knowledge of how the subject associates different concepts in memory. The authors did their best (within the constraints of limited recording time) to test images that might be conceptually related. In one tantalizing example, a neuron responded to both Jennifer Aniston and Lisa Kudrow, her co-star on the television show *Friends*. What seems to be a sparse representation in visual space may be a distributed representation in sitcom space! In another example, a neuron responded to two unrelated stimuli commonly used by Quiroga *et al.* — pictures of Jennifer Aniston with Brad Pitt and pictures of the Sydney Opera House. This could reflect a new memory association produced by the close temporal proximity of these stimuli during the recording sessions, consistent with similar phenomena observed in monkey temporal cortex¹⁵.

Thus, Quiroga and colleagues' findings may say less about visual representation as such than they do about memory representation and how it relates to visual inputs. Quiroga *et al.* have shown that, at or near the end of the transformation from visual information about

object structure to memory-related conceptual information about object identity, the neural representation seems extremely sparse and invariant in the visual domain. As the authors note, these are predictable characteristics of an abstract, memory-based representation. But I doubt that anyone would have predicted such striking confirmation at the level of individual neurons.

Charles E. Connor is in the Department of Neuroscience and the Zanvyl Krieger Mind/Brain Institute, Johns Hopkins University, Baltimore, Maryland 21218, USA.

e-mail: connor@jhu.edu

1. Rose, D. *Perception* **25**, 881–886 (1996).
2. Barlow, H. B. *Perception* **1**, 371–394 (1972).
3. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. *Nature* **435**, 1102–1107 (2005).
4. Pasupathy, A. & Connor, C. E. *Nature Neurosci.* **5**, 1332–1338 (2002).
5. Brincat, S. L. & Connor, C. E. *Nature Neurosci.* **7**, 880–886 (2004).
6. Young, M. P. & Yamane, S. *Science* **256**, 1327–1331 (1992).
7. Olshausen, B. A. & Field, D. J. *Nature* **381**, 607–609 (1996).
8. Vinje, W. E. & Gallant, J. L. *Science* **287**, 1273–1276 (2000).
9. Biederman, I. *Psychol. Rev.* **94**, 115–147 (1987).
10. Marr, D. & Nishihara, H. K. *Proc. R. Soc. Lond. B* **200**, 269–294 (1978).
11. Booth, M. C. & Rolls, E. T. *Cereb. Cortex* **8**, 510–523 (1998).
12. Bulthoff, H. H., Edelman, S. Y. & Tarr, M. J. *Cereb. Cortex* **5**, 247–260 (1995).
13. Vetter, T., Hurlbert, A. & Poggio, T. *Cereb. Cortex* **5**, 261–269 (1995).
14. Logothetis, N. K. & Pauls, J. *Cereb. Cortex* **5**, 270–288 (1995).
15. Sakai, K. & Miyashita, Y. *Nature* **354**, 152–155 (1991).

EARTH SCIENCE

New Madrid in motion

Martitia P. Tuttle

A new network of geodetic field stations has greatly improved monitoring of relative motion across a seismic zone in the central United States. It seems that rapid deformation is occurring across this fault system.

The New Madrid seismic zone lies 50–200 km from Memphis, Tennessee, and was the site of devastating earthquakes in 1811 and 1812. These earthquakes included three mainshocks and many aftershocks, with the largest earthquake having an estimated^{1,2} magnitude of 7.4–8.1. Historically, New Madrid has been the most seismically active region in central and eastern North America — what hazard might it pose today?

This question has been the subject of vigorous debate in the Earth science and earthquake engineering communities^{3,4}. The report by Smalley *et al.* (page 1088 of this issue)⁵ will enlighten that debate. From high-precision Global Positioning System (GPS) measurements, made with a newly installed network of field stations, they conclude that the New Madrid seismic zone is rapidly deforming at rates of the same order of magnitude as those at the boundaries of tectonic plates. This result

contradicts earlier estimates of low rates of deformation or strain accumulation⁶, but is consistent with geological evidence for the occurrence of repeated 1811–1812-type (New Madrid) events in the past 2,000 years^{7,8}.

During the past 12 years, geologists found a record of New Madrid events in the form of earthquake-related features, known as sand blows (Fig. 1, overleaf). The sand blows formed as a result of liquefaction, a process by which water-saturated sandy sediment below the surface is liquefied and vented on the ground in response to strong earthquake shaking. Detailed study of hundreds of sand blows, some of which are associated with Native American archaeological sites, led to the interpretation that they formed during three, possibly four, New Madrid events of magnitude 7.6 or greater in the past 2,000 years⁸.

In the 1990s, geophysicists undertook GPS measurements using a network of field

NATURE

50 YEARS AGO

"Personal Factors in Accident Proneness." Dr. J. A. Smiley... has made full use of his position as medical adviser to an aircraft-manufacturing company to study the accident histories of 6,450 men, and to examine in detail 87 men classified as accident prone... His thesis may briefly be stated — accident-prone individuals are usually emotionally disturbed, with associated hypothalamic misfunction which, it is tentatively suggested, produces minor imbalance of adrenalin and acetylcholine with concomitant behaviour disturbance... [they] also show 'anxiety' sweating in interview, albumin in the urine specimens collected during medical examination, a seven-fold increase in peptic ulcer incidence and a more than four-fold increase in incidence of other medical symptoms... The problem remains, however, whether these men may adequately be described as accident prone... the main conclusion to be drawn is that proneness to report minor injury can be added to the list of other known clinical signs of emotional disturbance.

From *Nature* 25 June 1955.

100 YEARS AGO

Prof. E. Wiedemann, of Erlangen, sends us a short statement of observations described in his work on electric discharges... He agrees with Mr. Jervis-Smith as to the action of ozone, and advises persons who work for a long while with influence machines not to have these machines situated in the working room. "Ozone belongs to the poisonous gases, and is the more dangerous, since the injurious effects are not manifest at the time; on the contrary, breathing the gas produces at first a feeling of exhilaration, but afterwards it has a depressing effect on the nervous system... During my observations I have suffered somewhat severely from nervous disturbance (hyperesthesia of the feet) due to breathing ozone. These lasted for one or two years. Moreover, I always experience discomfort after performing experiments in my lectures on Tesla discharges."

From *Nature* 22 June 1905.

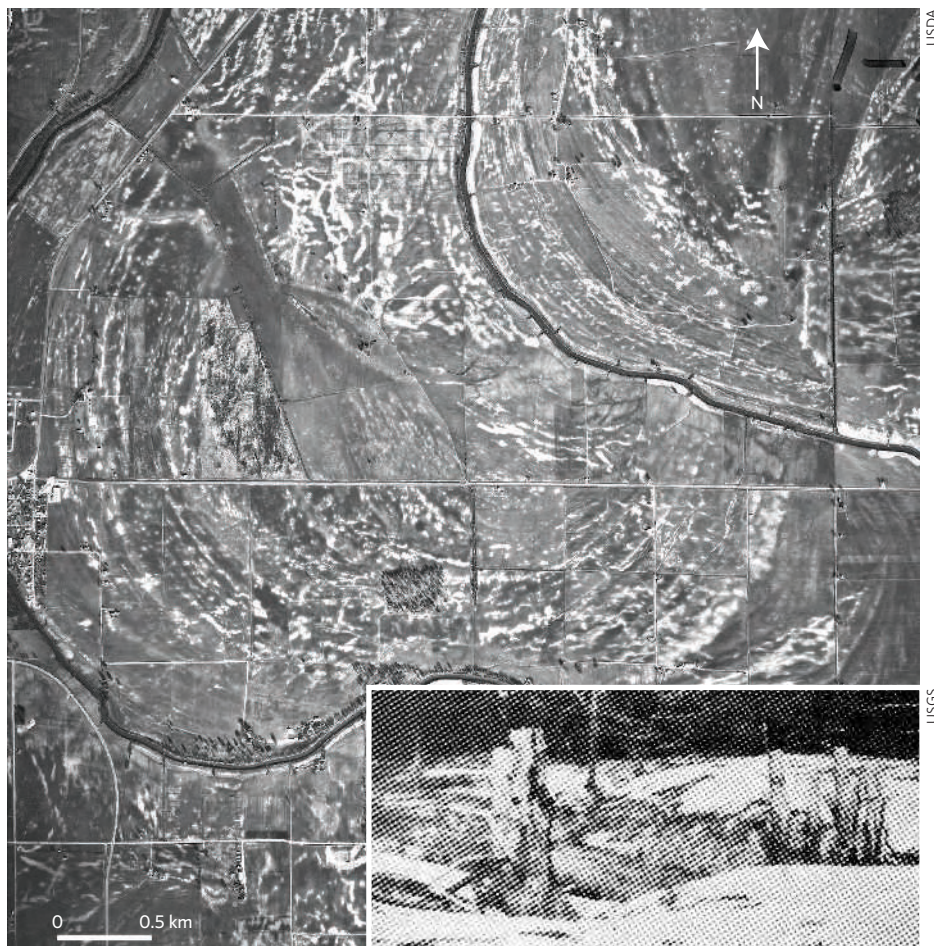


Figure 1 | Earthquake evidence. This aerial photograph, taken in 1964, shows light-coloured sand blows near the Little River in northeastern Arkansas. The inset is a ground view, taken about 100 years ago, of trees killed by the sand deposits. Some of the sand blows were produced by the New Madrid earthquakes of 1811–12; others were formed in prehistoric times. Smalley and colleagues' analyses⁵ are consistent with the finding of fairly frequently repeated New Madrid events surmised from this geological record.

stations spanning the New Madrid region to ascertain the rate at which the seismic zone is deforming in response to tectonic forces⁶. Measurements were collected for several days in 1991, 1993 and 1997, the upshot being estimated relative motion across the seismic zone of 1.4 mm yr^{-1} with uncertainties of $\pm 3 \text{ mm yr}^{-1}$. These motions were interpreted to be indistinguishable from zero, and therefore indicative of low rates of strain accumulation. Given that earthquake frequency is related to the build-up and release of strain energy, it was concluded that the New Madrid seismic zone produces either magnitude 8 earthquakes every 5,000–10,000 years or magnitude 7 earthquakes every 1,000 years⁶. This finding differed from that of the geological studies.

In the late 1990s, a network of permanent GPS stations was installed in the New Madrid region. The new network included many improvements; for example, stations were located close to and on both sides of major New Madrid faults, and strong H-beams were used that are less susceptible to non-tectonic movements than the 1-inch-diameter steel rods used in the previous network⁵. Because

the new stations are permanent and collect data continuously, the repeated setting up of field stations, which introduced measurement errors in the previous studies, could be avoided.

Smalley *et al.*⁵ have analysed four years of continuous measurements from the new network. They calculate relative motions across the seismic zone that are similar ($1\text{--}2.7 \text{ mm yr}^{-1}$) to those measured during the 1990s but with much smaller uncertainties — at most 25% of those of the previous studies. Smalley *et al.* point out that in the earlier GPS data the tectonic signal was lost in the noise, and interpret their results to indicate high rates of strain in the New Madrid seismic zone.

They also find relative motions across the seismic zone that are consistent with expected fault movements as inferred from present-day seismicity⁹ and recent fault studies⁷. For example, relative motion indicates that bedrock slips over itself along a major northwest-oriented fault, known as the Reelfoot thrust fault, that is inclined towards the southwest (see Fig. 2 on page 1089). The new findings are persuasive because they help to explain the geological observations of frequent New

Madrid earthquakes, and they make sense in terms of the active faulting in the region.

One of the most interesting results is that motions in the surrounding region are low compared with motion in the seismic zone itself. This unusual behaviour differs from that at plate boundaries, raising questions about the driving forces and earthquake processes within plates. Post-seismic afterslip — a process by which fault displacements at depths of several kilometres are expressed at the surface for a period of time following an earthquake¹⁰ — seems a reasonable explanation for the regional pattern of motions. However, there is currently insufficient information about the physical properties of the Earth in the New Madrid region to test this and competing models.

Smalley and colleagues' results are consistent with the findings of geological studies that the seismic zone produced earthquakes about every 500 years of magnitude 7.6 or greater. As such, they provide scientific justification for the adoption of stricter earthquake provisions in the building codes for Memphis and other cities in the central

United States⁴. Looking ahead, installation of additional field stations close to known faults would help to define their extent and further quantify their strain rates. One of the most daunting challenges will be to develop and test models that can explain how such large and frequent earthquakes are produced in the New Madrid region, and to see if the models also apply to other intraplate regions. ■

Martitia P. Tuttle is at M. Tuttle & Associates, 128 Tibbetts Lane, Georgetown, Maine 04548, USA. e-mail: mptuttle@earthlink.net

1. Johnston, A. C. *Geophys. J. Int.* **126**, 314–344 (1996).
2. Hough, S. E., Armbruster, J. G., Seeber, L. & Hough, J. F. *J. Geophys. Res.* **105**, 23839–23864 (2000).
3. Stein, S., Tomasello, J. & Newman, A. *Eos Trans. AGU* **84**, 177184–177185 (2003).
4. Frankel, A. *Seismol. Res. Lett.* **75**, 575–586 (2004).
5. Smalley, R. Jr, Ellis, M. A., Paul, J. & Van Arsdale, R. B. *Nature* **435**, 1088–1090 (2005).
6. Newman, A. S. *et al. Science* **284**, 619–621 (1999).
7. Kelson, K. I. *et al. J. Geophys. Res.* **101**, 6151–6170 (1996).
8. Tuttle, M. P. *et al. Bull. Seismol. Soc. Am.* **92**, 2080–2089 (2002).
9. Chiu, J. M., Johnston, A. C. & Yang, Y. T. *Seismol. Res. Lett.* **63**, 375–393 (1992).
10. Marone, C. *Annu. Rev. Earth Planet. Sci.* **26**, 643–696 (1998).

EVOLUTIONARY BIOLOGY

Island of the clones

Thomas N. Sherratt and Christopher D. Beatty

The discovery of an all-female population of damselflies in the Azores archipelago provides a novelty for entomologists. It also highlights the unique selection pressures faced by species that colonize islands.

Tucked away in the journal *Odonatologica* comes a paper by Cordero Rivera and colleagues¹ that will surprise many entomologists, and will exercise biologists studying evolution on islands and the mechanisms of sex determination. Cordero Rivera *et al.* have discovered that a species of damselfly on the Azores reproduces parthenogenetically (Fig. 1). This form of reproduction, in which females produce eggs that develop without fertilization by males², has been recorded in

almost all insect groups. But until now it was not known to occur in any natural populations of damselflies or dragonflies (the Odonata)³.

The Azores archipelago lies 1,500 km from the coast of Europe. Inspired by a report⁴ that only females of the damselfly *Ischnura hastata* had ever been found there, Cordero Rivera and his team visited 15 localities on six of the islands. Although more than 330 adult specimens of *I. hastata* were examined, none of them was male. To test whether the species



Figure 1 | Reproduction without fertilization in a damselfly. A female *Ischnura hastata* lays eggs in a pond on the island of Pico, Azores. (Courtesy of A. Cordero Rivera, Univ. Vigo.)

was parthenogenetic, a sample of larvae was reared to adulthood in the laboratory — more than 1,900 females were produced over nine generations, but no males.

Ischnura hastata is common in North and South America, yet it occurs in these regions as a classically sexual species with both males and females. The concept of 'geographic parthenogenesis'⁵ proposes that the parthenogenetic forms of a species are more likely to occur in certain areas — such as higher latitudes and altitudes, and on islands — because of the different selection pressures that organisms face under these conditions^{6,7}. One possibility, therefore, is that certain damselfly species can include both sexual and parthenogenetic forms, and that on arriving on a remote island it is the parthenogenetic form that is favoured, at least initially, owing to the difficulty of finding mates.

One might wonder why standard sexual reproduction does not kick in once the population builds up in size, but perhaps local conditions continue to favour parthenogenesis. Indeed, *I. hastata* frequents temporary or recently established habitats⁴, and Cordero Rivera *et al.* note that there is anecdotal evidence of local extinctions of pond populations. Furthermore, chance may play a role in the establishment and maintenance of parthenogenesis: *I. hastata* is also found on the Galapagos Islands, but the population contains both males and females⁸.

In at least some odonates, there may be a degree of predisposition to parthenogenesis; for example, there is evidence that unfertilized eggs of the dragonfly *Stylurus ocellatus* can be artificially induced to develop⁹. Moreover, certain parasites that are inherited only in the female line can manipulate their insect host into producing predominantly (or only) female offspring¹⁰. Cordero Rivera and colleagues are testing whether any microbial agents are responsible for driving the absence of males in *I. hastata*, but they have ruled out one potential bacterial parasite, *Wolbachia*, which infects a range of other insect groups¹⁰. If parthenogenesis in *I. hastata* is parasite mediated, then the microbial agent might have had a beneficial effect on its host in the initial phases of colonization, allowing individuals to reproduce without mates.

There have also been intriguing accounts of other damselfly species on remote archipelagos. In particular, on the islands of Fiji, it seems that females of the damselfly *Nesobasis rufostigma* actively defend territories over aquatic habitats, whereas the males, which are infrequently encountered, reside some distance from the stream¹¹. This phenomenon has been dubbed 'sex-role reversal'¹¹ and, if confirmed, would be the first example in an odonate. If males are in short supply, then this unusual mating system might be explained by female competition for access to males¹². Furthermore, males of two rarer Fijian damselflies (*N. flavostigma* and *N. caerulea*) have

never been found, raising the possibility that parthenogenesis occurs in these species.

Clearly, the evolutionary reasons why parthenogenesis is maintained in *I. hastata* remain largely unsolved. Entomologists, and those interested in island biology and parasite-mediated sex determination, have a new case for investigation. ■

Thomas N. Sherratt and Christopher D. Beatty are in the Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada.
e-mail: sherratt@ccs.carleton.ca

1. Cordero Rivera, A., Lorenzo Carballa, M. O., Utzeri, C. & Vieira, V. *Odonatologica* **34**, 1–9 (2005).
2. Hughes, R. N. *A Functional Biology of Clonal Animals* (Chapman & Hall, London, 1989).
3. Corbet, P. S. *Dragonflies: Behaviour and Ecology of Odonata* (Harley Books, Essex, 1999).
4. Belle, J. & Van Tol, J. *Tijdschr. Ent.* **133**, 43–147 (1990).
5. Vandel, A. *Bull. Biol. France Belg.* **62**, 164–281 (1928).
6. Cuellar, O. *Science* **197**, 837–843 (1977).
7. Glesener, R. R. & Tilman, D. *Am. Nat.* **112**, 659–673 (1978).
8. Turner, P. E. *Pan. Pac. Ent.* **43**, 285–291 (1967).
9. Watanabe, Y., Yokota, H., Kato, K. & Hatakeyama, M. *Proc. Arthropod. Embryol. Soc. Jpn* **34**, 31–32 (1999).
10. O'Neill, S. L., Hoffmann, A. A. & Werren, J. H. *Influential Passengers* (Oxford Univ. Press, 1997).
11. Donnelly, T. W. *NZ J. Zool.* **17**, 87–117 (1990).
12. Jiggins, F. M. et al. *Proc. R. Soc. Lond. B* **267**, 69–73 (2000).

CONSCIOUSNESS

Crick and the claustrum

Charles F. Stevens

Francis Crick believed that, in biology, structure is the natural path to understanding function. In his later career, he applied this dictum to the study of consciousness.

Pretty much everyone is interested in the big questions about the brain, and the biggest big question is: what is consciousness? Just as historically the vitalists could not imagine how life can be explained by just physics and chemistry — they believed that a non-physical 'life force' had to be involved — the dualists of today cannot believe our experience of the

feeling of love or the redness of red could arise just through nerve impulses in a bunch of brain cells. Although everyone who enters the field of neuroscience starts with an interest in the big questions, we soon settle into much smaller questions that we can see how to answer with the tools of modern biology. Questions about consciousness were therefore

mostly left to philosophers and kooks, and no respectable neuroscientist would even have considered working on such a problem — until Francis Crick, that is.

After he and James Watson solved one of biology's really big problems, the mechanism of inheritance, Crick moved to neuroscience and set himself the task of answering that field's biggest question. Working closely with Christof Koch, Crick made the study of consciousness respectable and, directly and indirectly, had a profound influence on all of neuroscience and on the types of questions that are considered acceptable to study. Crick's final paper, written with Koch, has just been published in *Philosophical Transactions of the Royal Society of London* (doi: 10.1098/rstb.2005.1661) and it proposes that an obscure part of the brain, the claustrum, may be involved in consciousness. Crick was working on this paper literally on his deathbed, and Koch has put the finishing touches on it for publication.

How can a scientist think about consciousness? Crick's approach had two parts. The first was to identify what properties of consciousness had to be explained, and the second was to find brain structures that might account for those properties. Crick and Koch note that a key feature of our conscious experiences is that all of the components are integrated into a unified whole: how a rose looks, smells and feels are bound together with our emotional experience of it. Because these different aspects of experience are related to neuronal

ANIMAL BEHAVIOUR

Congo's art

The decisive brush strokes are not the most notable feature of this painting, nor the powerful colour combinations. It is the artist and the artist's mentor — Congo the chimpanzee and Desmond Morris, respectively — that are the main points of interest. The painting, along with two others of Congo's, came up for auction at Bonhams, London, earlier this week.

Morris trained as an ethologist and has long been a painter himself. In the 1950s, he was the host of the television series *Zootime*, and it was here that Congo came to public attention. This picture was produced by the chimpanzee when he was three years old.

Congo was neither the first nor the last of ape artists, and his talent remains a question for experts in the disparate fields of art appreciation and animal behaviour. But his celebrity status has undoubtedly made his oeuvre more collectable.

Tim Lincoln



BONHAMS

processing in distinct and often widely separated brain circuits (those responsible for vision, olfaction, somatic sensation, together with the amygdala and other centres involved in emotion), this unification of experiential components implies some sort of coordination between different brain areas. In their survey of various notions about consciousness, Crick and Koch observe that a common thread through all of the thinking about consciousness is the recognition of a need to bind together information from many separate parts of the brain.

In the paper Crick writes, "In biology, if seeking to understand function, it is usually a good idea to study structure". And thus he takes a fundamentally structural approach to consciousness: what brain regions, he asks, have properties that would suit them for the information gathering and analysis that is at the heart of the conscious experience? I know from conversations with Crick that he had a very strong hunch, one that bordered on a conviction, that the structure underlying consciousness is the claustrum.

What is the claustrum, and why pick on it as a key for understanding consciousness? The claustrum is a thin sheet of grey matter that resides parallel to and below part of the cortex (the cortex is the grey matter covering of the brain that carries out the computations involved in feeling, seeing, hearing, language and deciding what to do). The claustrum is present in all mammals, but it has been little studied and its function is not known. What is known, however, is that there are two-way connections between the claustrum and most, if not all, parts of the cortex as well as subcortical structures involved in emotion.

So the claustrum is not just a sort of shadow of the cortex, but rather a neural circuit with overlapping inputs from various cortical regions and outputs back to cortex. Because of its widespread connections, Crick and Koch liken the claustrum to the conductor of an orchestra, who is responsible for binding the performances by individual musicians into an integrated whole that can be much more than the sum of the parts. The neuroanatomical connections of the claustrum, then, just match with the 'conductor' required to bind together the various disparate components of the conscious experience represented in many different brain regions.

Crick told me that one of his main purposes in this paper was to encourage new studies of the claustrum, and had he lived longer, he would have liked to start a centre for investigating the claustrum, where neuroanatomical, electrophysiological and novel molecular biological approaches to the claustrum could be combined. Some of these ideas for studying the claustrum, like using molecular biological methods to specifically disrupt claustral function, are sketched in this paper.

Not everyone will buy the Crick and Koch idea that the claustrum is the seat of con-

sciousness. For example, the fact that all mammals have a claustrum could be an argument against the proposal for those who cannot imagine consciousness without language and high-level symbolic reasoning. And I expect others will be sceptical on other grounds. Nevertheless, the proposal is an interesting and challenging one, from a scientific giant, and I

believe every scientist will be fascinated to see how one of the greatest biologists attacked such a difficult problem. ■

Charles F. Stevens is at the Molecular Neurobiology Laboratory, Salk Institute, 10010 North Torrey Pines Road, La Jolla, California 92037, USA.
e-mail: stevens@salk.edu

GRANULAR MATTER

A tale of tails

Martin van Hecke

Granular materials such as sand can either be jammed and rigid, or yield and flow. Puzzling changes in the forces between the grains deepen the mystery surrounding this basic, but poorly understood, transition.

Why does sand mimic a solid when we walk on it but emulate a fluid when it is in an hourglass? Why does salt flow only when its shaker is tipped far enough? These simple questions are without a clear answer, and have in recent years inspired investigations into what exactly happens when otherwise jammed¹ granular media lose rigidity and yield.

Two papers^{2,3} in this issue suggest that the key to jamming and yielding are the networks formed by the contact forces between individual granular particles (Fig. 1a). On page 1075, Corwin, Jaeger and Nagel² report that jammed and yielded glass beads differ in the statistics of their force networks, and, on page 1079, Majmudar and Behringer³ investigate the statistics and structure of force networks for jammed systems under compression or shear forces. The two experiments employ photoelastic materials (which rotate the polarization of light depending on stress, thus visualizing forces when viewed between polarizers) to obtain the grain-wall and grain-grain forces, respectively.

In Corwin and colleagues' experiment², a slowly rotating plunger exerted a constant,

shearing force on glass beads filling a cylindrical container. The motion of the bottom layer of beads revealed that beads near the sides flowed past each other, whereas those near the centre remained jammed and rotated as a solid block. Corwin and colleagues measured the forces on a photoelastic bottom plate, and express the probability of a grain exerting a force of a certain magnitude in a distribution such as that in Fig. 1c. Their central finding is that a change in the 'tail' of this distribution (characterizing the particles carrying the largest forces) signals the point at which the system jams: jammed grains produce tails with an exponential fall-off, whereas yielded grains produce much steeper tails. Thus flowing grains avoid large forces more effectively than those that are jammed.

Surprisingly, the yielded force distributions are rate-independent (that is, they do not vary with flow speed), and can be characterized by describing the flowing beads as if they form an ordinary liquid (at a constant temperature). It is usually tacitly assumed that these slow, rate-independent granular flows are quasi-static, so that a snapshot of a flowing

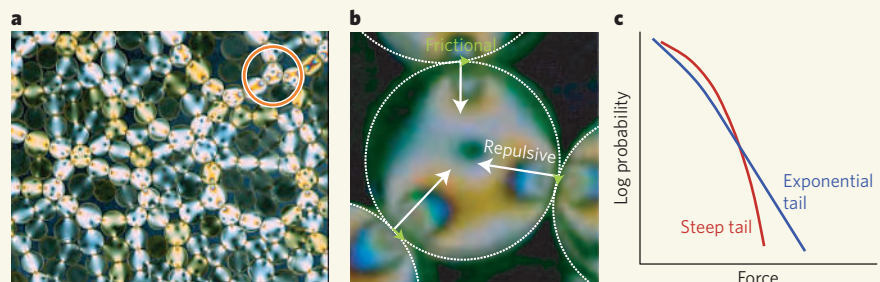


Figure 1 | Fluctuating forces. **a**, A force network, typical of granular media, revealed in a layer of photoelastic discs. The bright discs are experiencing the largest forces, and appear to align in 'force-chains'. **b**, An enlargement of the single particle indicated in (a) reveals a complex pattern of bright and dark bands, called fringes. From these, Majmudar and Behringer³ measured the repulsive normal (white) and frictional tangential (green) contact forces inside force networks. **c**, Probability distributions for the contact forces in granular media. Corwin, Jaeger and Nagel² relate jammed systems to the blue curve, which has an exponential tail, and yielded systems to the red curve, which has a much steeper tail.

Box 1 Packing problems

When spherical grains in a container are shaken down, at the densest possible packing they fill a volume fraction of around 64%. This is 'random close packing' — a notoriously controversial concept⁹, as regular periodic packings (similar to how oranges are packed in your grocery store) reach higher densities of 74%. Allowing small regular regions in disordered packings thus can increase the density beyond random close packing: 'random' and 'close' represent opposing trends⁹. (Incidentally, non-spherical grains, such as M&Ms, also pack more densely than 64%¹¹). Even more elusive is random loose packing¹⁰, which can be achieved by immersing spheres in a neutrally buoyant fluid and letting them settle gently, creating very fragile packings at volume fractions around 55%. Packing and jamming are related: soft, frictionless spheres jam, in the absence of shear, at a density precisely given by random close packing. Perhaps random loose packing might be defined, similarly, as the density where frictional spheres jam. **M.v.H.**

system cannot be distinguished from a stationary — jammed — state. Corwin and colleagues' experiment show that this assumption is incorrect — with the difference hidden in the force distributions.

What would happen if the rotating top disc of Corwin and colleagues' experiment were gradually stopped? The forces could simply freeze — but that would give yielded forces for jammed grains. The forces of the sheared grains could relax to a jammed distribution — but this would imply the breakdown of rate-independence. One solution to this conundrum could be the presence of additional characteristics such as packing density^{4,5} or anisotropy⁶ of the contact and force networks, which might differ between jammed and yielded grains. To uncover what goes on, we thus need to look inside granular media.

Majmudar and Behringer³ have done just that, investigating a granular material consisting of a layer of discs made of photoelastic plastic. The discs exhibit characteristic fringe patterns that encode the contact forces in the system, and, through the analysis of the resulting images, the authors obtained the first quantitative determination of force networks (Fig. 1b). They illustrate the power of this method by comparing a strongly jammed, uniformly compressed system with a weakly jammed system under pure shear (compressed in one direction and expanded in the other). Even though both systems are jammed and therefore static, the force networks of the two systems are very different: the sheared system exhibits strong anisotropies⁶, and 'force-chains' are much longer than is the case in a compressed system.

The tails of the force distributions established by Majmudar and Behringer³ hold a surprise: they change from steep for compressed, strongly jammed systems to exponential for sheared, weakly jammed systems. The tails determined by Corwin and col-

leagues², in contrast, become steep only when the grains flow, and are exponential for all jammed cases — irrespective of whether the system has experienced a shear stress. The two experiments probe somewhat different aspects of the force network, but, even so, their results are not easily reconciled.

Yielding by shear is the main mechanism by which grains are made to flow. Indeed, Osbourne Reynolds suggested more than 100 years ago a relation between packing density and yielding: flowing grains dilate⁴. These new experiments^{2,3} illustrate that force networks also play a crucial role by signalling stresses, jamming and yielding — in other words, the state of the granular system. The precise connection between packing geometry, force networks and jamming, however, is still a puzzle⁷.

Recently, theoretical progress has been made by considering simplified systems without friction or shear such as packings of deformable, frictionless particles. One of the most exciting findings is that the jamming–yielding transition in this system, which occurs when the confining pressure is lowered to zero, has many properties of a phase transition such as that which occurs between the solid and liquid states of matter. Near the critical point at which the transition occurs, the number of contacts reaches the minimal value allowed by mechanical stability, and the packing fraction approaches random close packing⁵ (Box 1): jamming, packing geometry and critical phenomena are thus connected. But what happens for systems that yield under shear? How do force networks fit into this picture? Are these ideas relevant to jamming and yielding of realistic frictional granular media?

Studies relating jamming to the packing

geometry for compressed frictional systems may start to bridge the gap between theory and experiment. The packing-densities of frictional granular media under low pressure span a wide range from random close packing to random loose packing^{5,8–10} (see Box 1). In the experiments conducted by Majmudar and Behringer³, frictional forces are small ('weakly mobilized'). But does this remain true for lower packing densities? Does random loose packing correspond to a maximal mobilization of friction⁸? Do frictional grains approach a critical point at random loose packing similar to frictionless grains at random close packing^{5,8}?

Simple questions of the behaviour of sand and salt lead to deep riddles and complex physics. Granular scientists, armed with marbles and plastic discs, are finding that some of these are now yielding to scrutiny.

Martin van Hecke is in the Kamerlingh Onnes Laboratory, Faculty of Mathematics and Natural Sciences, Leiden University, PO Box 9504, 2300 RA Leiden, The Netherlands.
e-mail: mvhecke@physics.leidenuniv.nl

1. Liu, A. J. & Nagel, S. R. *Nature* **396**, 21–22 (1998).
2. Corwin, E. I., Jaeger, H. M. & Nagel, S. R. *Nature* **435**, 1075–1078 (2005).
3. Majmudar, T. S. & Behringer, R. P. *Nature* **435**, 1079–1082 (2005).
4. Reynolds, O. *Phil. Mag.* **20**, 467–481 (1885).
5. O'Hern, C. S., Silbert, L. E., Liu, A. J. & Nagel, S. R. *Phys. Rev. E* **68**, 011306 (2003).
6. Snoeijer, J. H., Vlugt, T. J. H., van Hecke, M. & van Saarloos, W. *Phys. Rev. Lett.* **92**, 054302 (2004).
7. Daniels, K. E. & Behringer, R. P. *Phys. Rev. Lett.* **94**, 168001 (2005).
8. Makse, H. A., Johnson, D. L. & Schwartz, L. M. *Phys. Rev. Lett.* **84**, 4160–4163 (2000).
9. Torto, S., Truskett, T. M. & Debenetti, P. G. *Phys. Rev. Lett.* **84**, 2064–2067 (2000).
10. Onoda, G. Y. & Liniger, E. G. *Phys. Rev. Lett.* **64**, 2727–2730 (1990).
11. Donev, A. *et al. Science* **303**, 990–993 (2004).

NEUROSCIENCE

An intrusive chaperone

Anders S. Kristensen and Stephen F. Traynelis

Stargazin is best known for helping to ferry receptor proteins to the surface of neurons. The discovery that it has an unexpected additional role has widespread implications for the way that neurons talk to each other.

Cognition relies on the fast transmission of excitatory signals between neurons. To achieve this, neurotransmitters such as glutamate are released from one neuron into the synapse (the junction between neurons) where they are picked up by receptors on the opposing 'post-synaptic' cell. Glutamate receptors called AMPARs form ion channels embedded in the cell membrane that, upon binding of glutamate, open rapidly to allow cations to flood into the neuron — converting the chemical signal from the neurotransmitter into an elec-

trical pulse. In this issue, Tomita *et al.* (page 1052)¹ show that an accessory protein that helps to shuttle AMPARs into the membrane does double-duty to amplify the effectiveness with which glutamate opens the channel. AMPARs are among the most intensively studied of the neurotransmitter ion channels, so this discovery of an 'overlooked' accessory subunit is quite a surprise.

Tomita *et al.*¹ describe a functional analysis of the membrane-spanning protein Stargazin, which until recently was known only as a

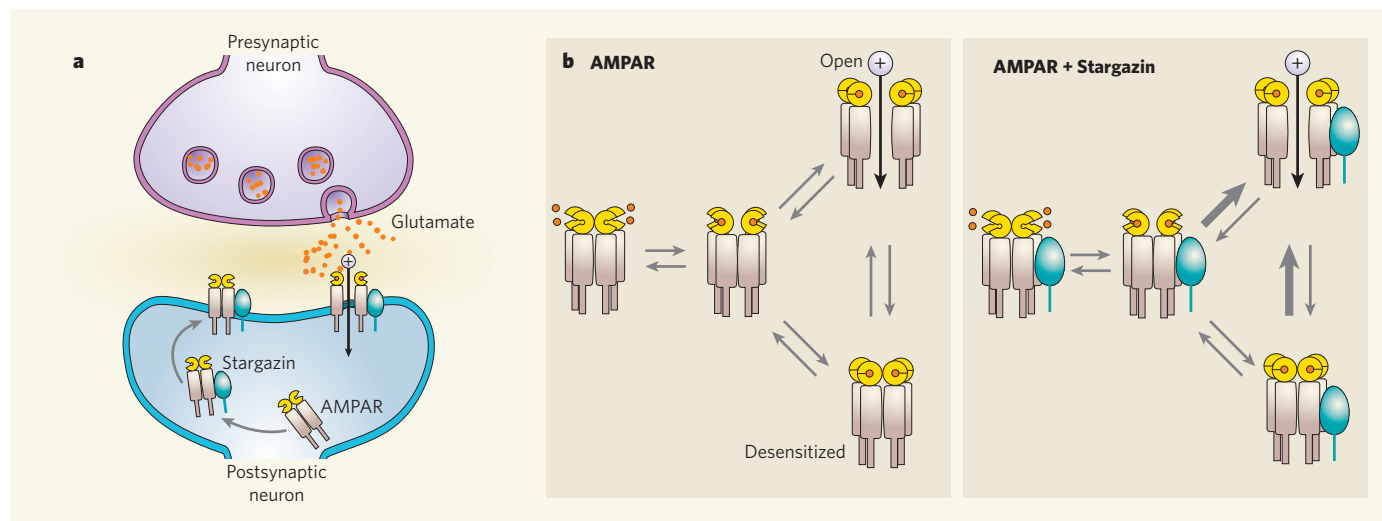


Figure 1 | Dual roles for Stargazin. **a**, Neuronal communication takes place at specialized junctions (synapses) formed between nerve terminals of transmitting (presynaptic) and receiving (postsynaptic) neurons. In excitatory synapses, glutamate is released from presynaptic membranes and binds to a group of glutamate receptors referred to as AMPARs. This leads to opening of ion channels and influx of cations, causing a brief electrical pulse. Stargazin engages AMPARs shortly after synthesis in the cell and influences transport of the receptors into the cell membrane. **b**, A model of AMPAR function whereby Stargazin controls receptor 'gating'. Following binding of glutamate, receptors can either undergo a conformational change into an open-channel state or a desensitized, closed state. Tomita *et al.*¹ suggest that Stargazin increases the rate by which AMPARs enter the open state, leading to enhanced ion flux during glutamate stimulation.

regulator that helped move AMPARs into the cell membrane. Their work shows that Stargazin can also control two key aspects of receptor function: the overall flux of ions through the water-filled AMPAR pore, and the ability of the receptor to function in the continued presence of glutamate, such as might occur during rapid neuronal firing. This finding, which has been corroborated by other recent studies^{2,3}, could have widespread implications for the way that neurons in the brain not only talk to each other, but how they increase or decrease their volume and change their pitch — features by which the brain may encode memory.

This is not the first time that Stargazin has surprised neuroscientists by having functions other than those initially proposed for it. Stargazin was originally discovered as a neuronal protein expressed from a gene mutated in the epileptic *Stargazer* mutant mouse⁴, and sequence analysis suggested it was a calcium-channel subunit. Then, it was the first integral membrane protein discovered to interact with AMPARs⁵. Subsequently, Stargazin and its family members were found to be necessary partners for AMPARs, both during transport to the neuronal membrane and the subsequent positioning of the receptors at synaptic sites. Highly coordinated spatiotemporal changes in the synaptic content of AMPARs are a principal mechanism by which the brain regulates synaptic strength. This activity-dependent regulation of excitatory signalling strength is considered a prime candidate for the biological mechanism underlying memory. Stargazin associates with AMPARs in an intracellular compartment near the cell membrane to chaperone their delivery to the neuronal surface, for example during changes in synaptic strength (Fig. 1).

Stargazin seems to remain in complex with AMPAR after delivery^{6,7}, indicating that most neuronal AMPARs are always in association with Stargazin or other family members. In this study, Tomita *et al.*¹ explore the potential contributions of Stargazin to AMPAR function once the receptor reaches the synaptic membrane (Fig. 1). They found that increases in the density of AMPARs at the cell surface could not entirely account for increases in the cell's functional responses to glutamate in the presence of Stargazin. An analysis of the kinetics of AMPAR function showed that the presence of Stargazin reduces the amount of time the receptor spends in the 'desensitized' state. In this state, glutamate remains tightly bound to the receptor, but the ion-conducting pore rarely opens. Moreover, the functional steps where the receptor closes its channel and glutamate unbinds also slowed in the presence of Stargazin (see also ref. 2).

These effects imply that Stargazin alters the ease with which the pore opens (referred to as gating) after glutamate binding (Fig. 1). Gating is defined as the conformational changes in the glutamate-bound receptor as energy gained from the interaction of glutamate with its binding site is used to open the ion channel. Recordings of currents from single channels indeed showed that in the presence of Stargazin, glutamate-bound AMPAR more frequently opens into high-conducting conformations — indicative of increased gating efficiency. Stargazin in native AMPARs should therefore enhance the efficiency with which glutamate released from nerve terminals can depolarize a postsynaptic neuron. This prediction was confirmed through experimentation with a mutated version of Stargazin that was unable to regulate AMPAR function but could facilitate AMPAR trafficking.

This new role for Stargazin raises obvious questions about previous kinetic and structural studies that used experimental systems lacking Stargazin or its siblings. In addition, it has become even more important to determine whether results obtained with recombinant receptors (produced from introduction of cloned DNA into cells) accurately reflect what happens with native neuronal receptors. There are also several complex questions that need to be answered about the accessory function of Stargazin. What factors control the size and nature of the AMPAR ion-conducting pore, and how does Stargazin regulate these factors? Can Stargazin's actions on permeation and gating themselves be regulated? Stargazin and other members of its family have discrete expression patterns in different regions of the central nervous system. Does such heterogeneity contribute to regional or developmental difference in AMPAR roles?

As always, much remains to be done to answer these and other questions. But for the moment, the uncovering of Stargazin's ability to regulate AMPAR function itself reminds us to expect surprises from even the best-studied systems. Who knows how many more crucial functions hide among proteins we think we understand? ■

Anders S. Kristensen and Stephen F. Traynelis are in the Department of Pharmacology, Emory University School of Medicine, Atlanta, Georgia 30322, USA.

1. Tomita, S. *et al.* *Nature* **435**, 1052–1058 (2005).
2. Priel, A. *et al.* *J. Neurosci.* **25**, 2682–2686 (2005).
3. Yamazaki, M. *et al.* *Neurosci. Res.* **50**, 369–374 (2004).
4. Letts, V. A. *et al.* *Nature Genet.* **19**, 340–347 (1998).
5. Chen, L. *et al.* *Nature* **408**, 936–943 (2000).
6. Nakagawa, T., Cheng, Y., Ramm, E., Sheng, M. & Walz, T. *Nature* **433**, 545–549 (2005).
7. Vandenberghe, W., Nicoll, R. A. & Bredt, D. S. *Proc. Natl Acad. Sci. USA* **102**, 485–490 (2005).

OBITUARY

Fred S. Rosen (1930–2005)

Immunologist, paediatrician and polymath

Fred Rosen, who died on 21 May — four days short of his seventy-fifth birthday and three days after a celebration to mark his retirement as president of the CBR Institute for Biomedical Research in Boston, Massachusetts — combined in his person a range of attributes unparalleled among his contemporaries. His scientific achievements, as one of the most cited authors in biomedical science, were allied to a legendary clinical acumen, and a warmth and authority that secured the trust and devotion of his patients.

Rosen passed most of his professional life at Harvard — at the Medical School, the Children's Hospital and the Center for Blood Research. This last he transformed, by dint of his formidable persuasive powers as a fund-raiser and his unerring eye for young talent, from an obscure relic into one of today's foremost centres of biomedical research. In his own field, that of hereditary immunodeficiencies — failures, commonly calamitous, at various sites in the labyrinths of the human immune system — he was undisputed monarch.

Rosen gave early proof of his scientific mettle. He took special pride in the discovery, while still a resident, of a disease (X-linked hyper-IgM syndrome), the manifestations of which had long baffled paediatricians. The antibodies, or immunoglobulins, in the blood are the first line of defence against infection, and one form, IgM, predominates at birth. Rosen had already established that what were known as 'natural' antibodies to preceding generations of immunologists were in fact IgM molecules, which are unable to cross the placenta: this is the basis of the susceptibility of babies to bacterial infections. IgM is normally supplanted soon after birth by gamma globulin, or IgG. Rosen showed that the lack of IgG in his patients arose from an absence of the population of cells (B cells) in which it is normally made. This in turn resulted from the primary defect, which lies in the complementary cells (T cells) that in the normal state activate the B cells; and finally Rosen and his colleagues identified the aberrant gene.

He went on to show that another lethal immune deficiency (X-linked agammaglobulinaemia) was characterized by a lack of all circulating B cells, and that the so-called common variable immunodeficiency could have a variety of causes; this revelation stimulated the search for unrecognized gene defects. All these discoveries shed new light on the workings of the normal immune system and redirected

much of the thinking in the field.

The complement cascade, a set of surveillance proteins circulating in the blood, is another part of the immunological armoury. These proteins act by attracting phagocytes, specialized cells that engulf and destroy bacteria or other interlopers. Rosen and his colleagues made a lasting impact on many aspects of this system — by the discovery, for example, of an inhibitor that lies at the root of hereditary malfunctions of one of the proteins (C1). Perhaps the most spectacular coup concerned another complement constituent, C3, a deficiency of which Rosen was able to link to a previously unfathomed disease in children. This proved the key to several other hereditary conditions, and further studies by Rosen and his associates eventually uncovered the chemical route by which C3 provokes phagocytosis.

Rosen's last great contribution centred on an anomalous molecule found on the surface of T cells in patients suffering from a severe malady called Wiskott–Aldrich syndrome. Its chemical nature led Rosen to divine that the anomaly might engender a disturbance in the cell's internal network of filaments governing shape and movement. So indeed it proved, and the discovery initiated a new departure in cell biology.

There was more, but what united all of Rosen's undertakings was an acute sense of purpose, for his concern throughout was to understand the genesis of diseases, and to devise methods of treatment. He promoted the now universal treatment of immune deficiency with intravenous IgG, and thereby saved many lives and much misery. He and his colleagues performed the first successful bone-marrow transplants for two of the most dire immune conditions. This was quintessential 'translational research', long before that dispiriting term was invented.

Albert Szent-Györgyi viewed the practice of science as "seeing what everybody else has seen, and thinking what nobody else has thought". This was Rosen's supreme gift, and his capacity to discern an interesting, important and, above all, soluble problem attracted a succession of talented and percipient colleagues, all eager to take up the challenges that he defined, and with all of whom he worked closely. Through everything, he treated, comforted and cheered his paediatric patients, for he had an especial rapport with the young, and he taught generations of students. His very appearance, that of an ample, genial teddy



CBR INST.

bear, radiated reassurance. His prodigious knowledge and sagacity made him a universal fount of information and good sense — the "primal Google", in the words of one of his acolytes.

Fred Rosen had friends in all parts of the world. He could converse, write and lecture in any of a half-dozen or more languages. He was especially drawn to Russia and its literature, music and culture. He was instrumental in setting up a paediatric leukaemia unit in St Petersburg. On a typical occasion, he returned from South Africa, having interviewed and collected blood samples from a family with a new type of complement anomaly. They were Afrikaaner farmers, he recounted, and he had been surprised to find that they spoke no English at all. How then could he communicate with them? He looked mildly surprised at such an inane question: why, you could speak Dutch to them, of course.

Rosen, in short, was a formidable polymath. He read omnivorously and retained it all. He knew more about antique furniture and silver than the dealers whom he patronized. He shopped at Fauchon and Hermès, had his shirts made in Jermyn Street and took tea among the dowagers at Fortnum & Mason. His friends were astonished that this archetypal urbanite should have built a graceful house on the island of Anguilla in the Caribbean, and have found a warm affinity with the local community. It was there that he decided his remains should rest.

Walter Gratzer and David G. Nathan

Walter Gratzer is at King's College, The Randall Division, Guy's Campus, New Hunt's House, London SE1 1UL, UK. David G. Nathan is at the Dana-Farber Cancer Institute, the Children's Hospital Boston and Harvard Medical School, Boston, Massachusetts 02115, USA. e-mails: wbg@helios.rai.kcl.ac.uk; David_Nathan@dfci.harvard.edu

BRIEF COMMUNICATIONS

Floater clustering in a standing wave

Capillarity effects drive hydrophilic or hydrophobic particles to congregate at specific points on a wave.

How do waves affect the distribution of small particles that float on water? Here we show that drifting small particles concentrate in either the nodes or antinodes of a standing wave, depending on whether they are hydrophilic or hydrophobic, as a result of a surface-tension effect that violates Archimedes' law of buoyancy. This clustering on waves may find practical application in particle separation and provides insight into the patchy distribution on water of, for example, plastic litter or oil slicks.

According to Archimedes' law, the mass of liquid displaced by an object is equal to the mass of that object. Small hydrophilic particles violate this law because capillarity (surface tension) pulls them deeper into the water, so the mass of the displaced liquid exceeds the particles' mass; small hydrophobic particles violate it because their mass is greater than that of the liquid they displace. Therefore when a hydrophilic/hydrophobic particle lies on a static, inclined liquid surface, there is an upwards/downwards force that is proportional to the surface slope: this can be demonstrated with hydrophilic and hydrophobic particles, which respectively climb up and slide down the meniscus (results not shown).

Standing wave patterns are characterized by certain points that undergo no vertical displacement (nodes). Fluid moves horizontally around the nodes, and midway between them are antinodes, where fluid moves purely vertically. Points on the fluid surface between a node and antinode move along inclined lines. Objects floating on such a wave should simply oscillate with the fluid, but an analysis of the interplay between gravity and capillarity shows that a net force acts on small particles floating between nodes and antinodes on a standing wave on a horizontal water surface (see supplementary information).

Let us compare the forces acting on a small particle in its high and low positions during the wave period (for calculations, see supplementary information). The first contribution to the net force is due to the fact that a particle placed between a node and antinode moves along a surface that is steeper nearer the node. In one half of the cycle, when the fluid surface is below the horizontal, the force acting on a particle is larger than in the other half of the cycle, when the surface is above the horizontal. As a result,

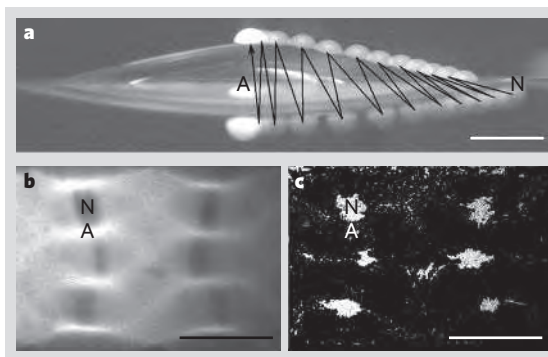


Figure 1 | Floaters in a standing wave. **a**, Stroboscopic picture (side view) of the drift of a hydrophobic teflon ball towards the point of maximum vertical displacement (antinode, A). Arrow shows the direction of the motion, which starts at the point of zero displacement (node, N). The line segments connect the centres of the ball positions at successive half-periods of the wave. **b**, **c**, Plan views of a standing wave with a shadowgraph (**b**), and showing hydrophilic particles clustering at its nodes (**c**). Scale bars, 5 mm.

there is a net force that pushes a hydrophobic particle towards the antinode, and a hydrophilic particle towards the node.

The second contribution to the force asymmetry is due to the vertical displacement of the floater. Because of inertia, the motion of a 'heavy' hydrophobic particle is retarded, whereas that of a 'light' hydrophilic particle is advanced relative to the motion of the fluid surface (as long as the wave frequency is less than the frequency of free particle oscillations — which is always the case when the wavelength is much larger than the particle size). The acceleration of the fluid surface is downwards when the fluid surface is above the horizontal, and upwards when it is below it. As a result, a hydrophobic particle is submerged to a greater depth at the lower position. Vertical displacement of a floater from its equilibrium position relative to the fluid surface produces an extra (restoring) force. This force pushes a hydrophobic particle up from its low position (and down from its high position), so that its horizontal component is directed towards an antinode. Conversely, a hydrophilic particle is submerged to a greater depth when in its higher position on the surface and the net force is towards a node.

Both these mechanisms are inertial (due to mass mismatch) and produce net forces in the same direction. The force is proportional to the square of the wave amplitude and so is a nonlinear effect.

We generated a standing wave by vertically oscillating a container¹ (see supplementary information for details of the experimental set-up). Floating hydrophobic particles attract each other, so we avoided clustering unrelated to waves by using a single hydrophobic teflon ball (density, 2.2 g cm⁻³; diameter, 1.6 mm). Figure 1a shows the motion of the ball, lit stroboscopically with twice the frequency of the standing wave (12.750 hertz). The stroboscope phase was adjusted to visualize only the instants of the maximal and minimal elevation of the water surface. The time taken to pass from a node to an antinode varies from 2 to 20 s, depending on the wave amplitude.

To observe the opposite direction of drift, and to study smaller particles, we used a suspension of hydrophilic hollow glass spheres of average size 30 micrometres and density 0.6 g cm⁻³. These particles submerge almost fully and readily form homogeneous suspensions in the absence of waves. The cell was illuminated from below and the light was recorded by two cameras synchronized with the shaker oscillations. Figure 1b shows a shadowgraph image from a camera that receives light passed straight upwards so that the white lines correspond to the wave antinodes and the black spots are shadows from the particles. Figure 1c shows an image of the same area taken by a camera that is placed to receive only light scattered by the particles so that the white spots represent particle clusters. The hydrophilic particles are clearly clustered in the nodes of the standing wave, as predicted by our calculations.

Particles on fluid surfaces can be compared with clustering phenomena such as Chladni patterns, in which sand on a vibrating elastic solid forms nodal patterns^{2,3}. Separation can also occur in such patterns (coarse grains to nodes, fine powder to antinodes), although this is a collective phenomenon^{4,5} distinct from our single-particle effect. Recall also that dust gathers at the nodal points of velocity in a standing acoustic wave⁶ (in Kundt's tube) owing to the radiation pressure gradient⁷. However, for waves in fluids, pressure is constant along the surface, so the phenomenon we describe here is different.

G. Falkovich*, A. Weinberg*, P. Denissenko†, S. Lukashuk†

*Physics of Complex Systems, Weizmann

Institute of Science, Rehovot 76100, Israel
 ‡Fluid Dynamics Laboratory, University of Hull,
 East Yorkshire HU6 7RX, UK

1. Faraday, M. *Phil. Trans. R. Soc. Lond.* **121**, 319–340 (1831).
2. Chladni, E. F. *Entdeckungen über die Theorie des Klanges* (Wittenberg, 1787).
3. Waller, M. *Chladni Figures: A Study in Symmetry* (Bell, London, 1961).

4. Faraday, M. *Phil. Trans. R. Soc. Lond.* **121**, 299–318 (1831).
5. Thomas, B. & Squires A. M. *Phys. Rev. Lett.* **81**, 574–577 (1998).
6. Rayleigh, L. *Phil. Trans. R. Soc. Lond. A* **175**, 1 (1883).
7. King, L. V. *Proc. R. Soc. Lond. A* **147**, 212–216, (1935).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
 doi:10.1038/4351045a

BIODIVERSITY

Disease threat to European fish

The deliberate introduction of new species can have unexpected negative consequences^{1,2} and we show here how a recently introduced fish, the invasive Asian cyprinid *Pseudorasbora parva*, is causing increased mortality and totally inhibiting spawning in an already endangered native fish, the European cyprinid *Leucaspis delineatus*. This threat is caused by an infectious pathogen, a rosette-like intracellular eukaryotic parasite that is a deadly, non-specific agent. It is probably carried by healthy Asian fish, and could decrease fish biodiversity in Europe, as well as having implications for commercial aquaculture.

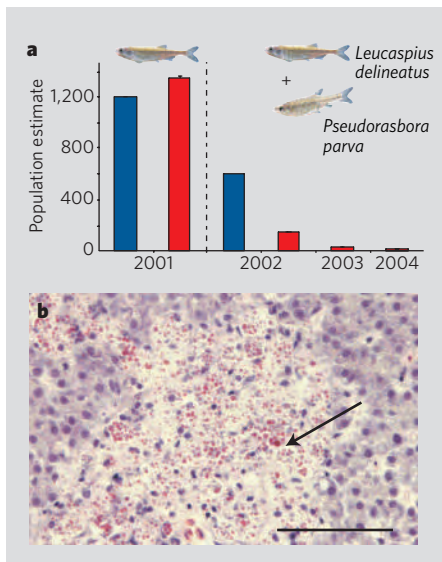


Figure 1 | Decline of *Leucaspis delineatus* population in a large natural pond after the introduction of *Pseudorasbora parva* and its associated pathogen, *Sphaerothecum destruens*. **a**, The exact maximum likelihood estimates (red bars, showing s.e.m.) for *L. delineatus* populations in the autumn are shown for four years. Starting populations in April of 2001 and of 2002 are shown as blue bars: these represent 1,200 *L. delineatus* in 2001, and 600 *L. delineatus* present in a mixed population with 600 *P. parva* introduced in 2002. **b**, High-power light micrograph of a section of a *L. delineatus* liver. There is a focus of phagocytic cells (arrow) containing conspicuous eosinophilic rosette-like agents (pink). (Slide stained with haematoxylin and eosin; scale bar, 50 micrometres.)

The sunbleak, *L. delineatus*, is the only representative of its genus and the only nest-guarding fish among European cyprinids. Once widespread in Europe, in the past 40 years it has inexplicably declined³ and is now on the European list of threatened freshwater fishes³. By contrast, since its introduction in 1960 into Romanian ponds near the River Danube, the Asian topmouth gudgeon, *P. parva*, has spread rapidly throughout Europe⁴ and has locally coincided with *L. delineatus* extinction^{5,6}.

In laboratory experiments (for methods, see supplementary information), we found that the holding water of *P. parva* acted as an absolute inhibitor of spawning for *L. delineatus* (no eggs produced in *P. parva* water compared with $1,596 \pm 840$ in control, clean water), and caused a large increase in fish mortality ($69 \pm 3\%$ deaths in the treatment group, compared with $16 \pm 2\%$; $P < 0.05$, Mann-Whitney *U*-test; 4 experiments). These results were confirmed in a large natural outdoor pond, where *L. delineatus* populations declined by 96% over three spawning seasons (2002–04; Fig. 1a) after being mixed with *P. parva*, despite an increase of 13% in the year before *P. parva* arrived (2001; Fig. 1a). Spawning was totally inhibited in *L. delineatus* after *P. parva* was introduced.

We found that the decline in *L. delineatus* (caused by total inhibition of spawning, loss of body condition, and death) that resulted from sharing water with *P. parva* was caused by an infectious organism. Histological findings from moribund *L. delineatus* indicated extensive infection of visceral organs, including the reproductive tissues, with an obligate intracellular eukaryotic pathogen (Fig. 1b; see also supplementary information) similar to the lethal rosette agent *Sphaerothecum destruens*⁷ that infects Chinook salmon, *Oncorhynchus tshawytscha*, and Atlantic salmon, *Salmo salar*.

The presence of this pathogen in *L. delineatus* that had been exposed to holding water from *P. parva*, and its absence in the source population and in the control group ($n=20$), was confirmed by polymerase chain reaction amplification of its DNA using primers specific to a small segment of the ribosomal DNA of *S. destruens*. The prevalence of the rosette-

like agent in moribund or dead *L. delineatus* was 67% ($n=12$). The parasite was also detected in subclinical fish in the treatment group, but at a lower prevalence (28%; $n=32$). This level of infection is consistent with that reported in tissues of salmon exposed to *S. destruens*⁸.

Preliminary examination indicates that other cyprinids, such as the fathead minnow *Pimephales promelas*, are also susceptible to this pathogen, which causes effects identical to those in *L. delineatus* (prevalence, 20%; $n=5$). All *P. parva* specimens ($n=10$) tested for the rosette-like agent were negative: however, this is to be expected, given that pathogen concentrations in healthy carrier fish are very low and difficult to detect using conventional diagnostic tests⁹. Cohabitation studies are a recognized method for detecting carrier states for different fish pathogens^{9,10} and, as our results illustrate, they are currently the most reliable way to detect a healthy carrier.

Our results have three important implications. First, the most invasive fish species in Europe⁴ is a healthy host for a deadly, non-specific pathogen that could threaten aquaculture trade, including that of salmonids. Second, it is difficult to identify fish populations that are carriers of pathogens. Third, this pathogen could pose a threat to the conservation of European fish diversity.

Rodolphe E. Gozlan*, **Sophie St-Hilaire†**,
Stephen W. Feist‡, **Paul Martin‡**,
Michael L. Kent§

*Centre for Ecology & Hydrology,
 Winfrith Technology Centre, Dorchester,
 Dorset DT2 8ZD, UK
 e-mail: reg@ceh.ac.uk

†Department of Biological Sciences, Idaho State
 University, Pocatello, Idaho 83209, USA

‡CEFAS Weymouth Laboratory, Weymouth,
 Dorset DT4 8UB, UK

§Center for Fish Disease Research, Department of
 Microbiology, Oregon State University, Corvallis,
 Oregon 97331-3804, USA

1. Lodge, D. M. *Trends Ecol. Evol.* **8**, 133–137 (1993).
2. Van der Zanden, J. M., Casselman, J. M. & Rasmussen, J. B. *Nature* **401**, 464–467 (1999).
3. Lelek, A. *The Freshwater Fishes of Europe: Threatened Fishes of Europe* (AULA, Wiesbaden, 1987).
4. Gozlan, R. E., Pinder, A. C. & Shelley, J. J. *Fish Biol.* **61**, 298–300 (2002).
5. Giurca, R. & Angelescu, N. *Romania. Bul. Cerc. Pisc.* **30**, 99–109 (1971).
6. Wolfram-Wais, A., Wolfram, G., Auer, B., Miksch, E. & Hain, A. *Hydrobiologia* **409**, 123–129 (1999).
7. Arkush, K. D., Mendoza, L., Adkinson, M. A. & Hedrick, R. P. *J. Eukaryot. Microbiol.* **50**, 430–438 (2003).
8. Mendonca, H. L. & Arkush, K. D. *Dis. Aquat. Org.* **61**, 187–197 (2004).
9. St-Hilaire, S., Ribble, C., Traxler, G., Davies, T. & Kent, M. L. *Dis. Aquat. Org.* **46**, 173–179 (2001).
10. Gilad, O. et al. *Dis. Aquat. Org.* **48**, 101–108 (2002).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
 doi:10.1038/4351046a

BRIEF COMMUNICATIONS ARISING online
 ▶ www.nature.com/bca see Nature contents.

PLANT COMMUNITIES

Ecosystem stability in Inner Mongolia

Arising from: Bai, Y., Han, X., Wu, J., Chen, Z. & Li, L. *Nature* **431**, 181–184 (2004)

Bai *et al.*¹ suggest that in China's Inner Mongolia steppe, community-level stability arises from compensatory effects among the principal components at both the species and plant functional group (PFG) levels. By analysing a consistent 19-year data set (1980–98), we show here that their analysis of a 24-year field data set (1980–2003) is called into question by inconsistencies in sampling location and numbers after 1998; the authors' findings are further undermined because they do not distinguish temporal variation from spatial heterogeneity in analysing compensatory effects among species or PFGs. We believe that rigorous reanalysis is needed for a better understanding of grassland stability.

The 24-year biomass data used by Bai *et al.*¹ are inconsistent owing to changes in sampling area and numbers after 1998, which accounts

for the discrepancy with our reanalysis of a consistent 19-year data set covering the same period but ending in 1998 (Table 1). We also excluded some extraordinarily wet years (see Methods), as do Bai *et al.*¹, who assert that January–July precipitation was the primary climatic factor causing fluctuation of community biomass production. However, we found that correlations were greatest between, on the one hand, the community biomass (measured by live above-ground biomass (LAB), as in Bai *et al.*¹) on both 15 and 31 August, and on the other, precipitation during the plant growing season (15 April to 15 August) ($r = 0.682$ and 0.705 , $P < 0.001$). We also found that annual precipitation was significantly correlated both with LAB on 15 August and with the annual peak LAB ($r = 0.547$ and 0.556 , $P < 0.05$) in the *Leymus chinensis* community. We found

no significant correlations between LAB or peak LAB and precipitation in the *Stipa grandis* community ($P > 0.05$). Furthermore, we found no compensations between dominant and subdominant, or between subdominant and non-dominant, species and PFGs in the *L. chinensis* community. The compensatory effect between dominant and subdominant PFGs was not evident in either site (Table 1).

Second, even neglecting the change of sampling location and resampling of the 5 quadrats from the 20 quadrats in the period 1999–2003, Bai *et al.*¹ do not convincingly demonstrate the existence of specific compensatory effects on the community stability — these are generally measured by temporal variation^{2–5}. Bai *et al.*¹ analysed data on relative figures for LAB to demonstrate compensations between PFGs or species (Figs 2 and 3 in Bai *et al.*¹). However, the relative mass of one PFG or species in a community would inevitably rise (or fall) if the relative mass of the other PFGs or species fell (or rose), irrespective of whether true compensation exists between them.

Bai *et al.*¹ also illustrate the compensatory effects by showing the negative correlations between species or PFGs (Table 1 in Bai *et al.*¹), but they pooled data for all 120 quadrats (5 quadrats per year for 24 years), which included both the temporal variability (generated by the 24 years) and the spatial variability (generated by the 5 replicates and the change of sampling location after 1998). Spatial heterogeneity alone can produce significantly negative correlations between some species, especially between dominant and subdominant species in both communities (Table 2), that are even stronger than those that Bai *et al.*¹ find in the bulk data set. Similar results for dominant–subdominant PFGs were also observed in both sites ($r = -0.500$, $P = 0.025$ in *L. chinensis* site and $r = -0.901$, $P < 0.001$ in *S. grandis* site). Furthermore, correlation does not imply causation⁶. The small proportion of significantly negative correlations in Bai *et al.*¹ (10 in 231 species pairs and 15 in 120 species pairs in sites A and B, respectively) is almost exactly what would be expected by chance, with a criterion of $P < 0.05$. It is not robust evidence for specific compensation effects.

Shipping Wang*†, Haishan Niu‡, Xiaoyong Cui‡, Shu Jiang*, Yonghong Li*, Xiangming Xiao§, Jinzhi Wang*, Guojie Wang*, Dehua Huang*, Qihui Qi*, Zonggui Yang*

*Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

†Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, China

‡Graduate School, Chinese Academy of Sciences, Beijing 100049, China

Table 1 | Number of significant correlations in species and plant functional group levels.

Leymus chinensis community					Stipa grandis community			
		Tested pairs	Number of PCs	Number of NCs	D–SD	Tested pairs	Number of PCs	Number of NCs
Species	15 Aug	351	18	4	/	231	28	11
	31 Aug	351	16	12	/	231	18	5
PFG	15 Aug	28	2	0	/	28	0	1
	31 Aug	28	0	0	/	28	1	1

PC, positive correlation; NC, negative correlation; D–SD, correlation between dominant–subdominant species or PFGs; PFG, plant function group, as in Bai *et al.*¹. Forward slashes, no correlation ($P > 0.05$); dashes, significant negative correlation ($P < 0.05$).

Methods. Study sites have been described^{1,7–9}. Every year the sampling started on 1 May, ended on 30 September and was done at 15- or 16-day intervals. For the period of 1980–98, five 1 m × 1 m quadrats were sampled randomly within the 100 m × 100 m permanent site; however, since 1999 the field-sampling design for *Leymus chinensis* community was changed and 20 quadrats were sampled in a new site (60 m north of the original site). The new and old *L. chinensis* sites differ in plant community structure: for example, *L. chinensis*, *Stipa grandis*, *Achnatherum sibiricum* and *Axyris amaranthoides* accounted for 43.1%, 12.4%, 8.7% and 0%, respectively, of community above-ground biomass at the old site in 1993, and 2.2%, 36.0%, 14.9% and 10.7%, respectively, in the new site in 2003. Both 1993 and 2003 had comparable amounts of precipitation over January–July and during the plant-growing season (15 April to 15 August). Over the 1980–98 and 1999–2003 sampling periods, transects were not designed, nor were there five equally sized replicate blocks, as Bai *et al.* claim. Bai *et al.*¹ do not detail changes in sampling area and methods as they were not personally involved in data collection before 1998, neither do they describe how 5 out of the 20 samples collected each year after 1998 were selected. To retain consistency in the data set and avoid the effect of spatial heterogeneity, we used only a 19-year data set (1980–1998; 18 years in some cases because only total biomass was available for 15 August 1986 in the *L. chinensis* community and there were no data for 31 August 1998 in the *S. grandis* community). Meanwhile, the data of 20 quadrats in 2002 were used to analyse spatial variation in *L. chinensis* and *S. grandis* communities. All analyses were done as described¹. At the species level, only commonly found species (those present in more than 50% of total sampling quadrats or years) were included. To analyse correlations between live above-ground biomass and precipitation, we excluded the extremely wet years (January–July precipitation, 1990, 1991, 1992 and 1998 excluded; precipitation in plant growing season, 1981, 1990, 1992 and 1998 excluded; and annual precipitation, 1981, 1986, 1990, 1992 and 1998 excluded), as suggested in Bai *et al.*¹.

Table 2 | Significant correlation coefficients between species in 20 quadrats in each site in 2002.

Leymus chinensis community				Stipa grandis community	
Species pair†	Correlation coefficients (r)	Species pair	Correlation coefficients (r)	Species pair	Correlation coefficients (r)
1 × 9	−0.539*	1 × 14	0.462*	1 × 3	−0.893**
8 × 19	−0.520*	10 × 12	0.599**	3 × 5	−0.708**
1 × 17	−0.512*	10 × 13	0.558**	3 × 8	−0.559*
8 × 13	−0.503*	12 × 16	0.500*	3 × 7	−0.546*
1 × 2	−0.490*	13 × 19	0.667**	4 × 8	−0.521*
8 × 9	−0.450*	14 × 20	0.463*	5 × 9	0.510*
3 × 16	0.444*	5 × 16	0.447*	7 × 10	0.532*
7 × 18	0.446*	8 × 15	0.578**	5 × 10	0.603*

Single and double asterisks indicate significance at the 0.05 and 0.01 levels, respectively.

†Numbers indicate the rank of the species in relative above-ground biomass in each site.

Institute for the Study of Earth, Oceans and Space, University of New Hampshire, Durham, New Hampshire 03824, USA

1. Bai, Y., Han, X., Wu, J., Chen, Z. & Li, L. *Nature* **431**, 181–184 (2004).
2. Doak, D. F. *et al.* *Am. Nat.* **151**, 264–276 (1998).
3. Tilman, D. *Ecology* **77**, 350–363 (1996).
4. McNaughton, S. J. *Am. Nat.* **111**, 515–525 (1977).
5. McNaughton, S. J. *Ecol. Monogr.* **55**, 259–294 (1985).
6. Huston, M. C. & McBride, A. C. in *Biodiversity and Ecosystem*

Functioning: Synthesis and Perspectives (eds Loreau, M., Naeem, S. & Inchausti, P.) 47–60 (Oxford Univ. Press, Oxford, 2002).

7. Xiao, X. M., Wang, Y., Jiang, S., Ojima, D. S. & Bonham, C. D. *J. Arid Environ.* **31**, 283–299 (1995).
8. Xiao, X. M., Jiang, S., Wang, Y., Ojima, D. S. & Bonham, C. D. *Vegetatio* **123**, 1–12 (1996).
9. Chen, Z. Z. & Wang, S. P. (eds) *Typical Steppe Ecosystem of China* (Science, Beijing, 2000).

doi:10.1038/nature03862

PLANT COMMUNITIES

Ecosystem maturity and performance

Arising from: Bai, Y., Han, X., Wu, J., Chen, Z. & Li, L. *Nature* **431**, 181–184 (2004)

The effect of maturity, or successional stage, on ecosystem performance (measured as productivity or stability, for example) is important for both basic ecology and ecosystem management. On the basis of the results of a long-term study of two different plant communities at two sites in the Inner Mongolia grassland¹, Bai *et al.* claim that these communities simultaneously achieve high species richness, productivity and ecosystem stability at the late successional stage¹. However, I question their interpretation of the data and suggest that this claim is undermined by evidence from other empirical and theoretical studies.

Bai *et al.*¹ present data on above-ground biomass of the total community and on various plant groups, but they provide no time series to indicate how diversity, productivity and stability might have changed during succession.

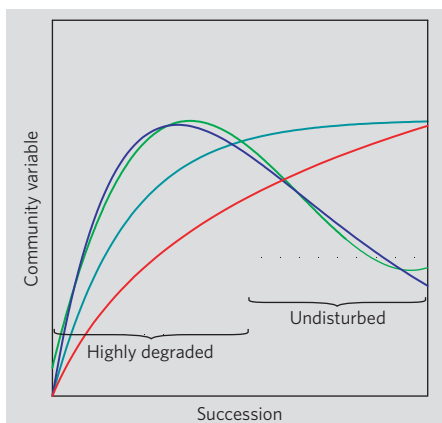


Figure 1 | Changes during succession. Model of temporal changes in diversity (species richness; green curve), productivity (blue), biomass (turquoise) and stability (red) during succession, based on an extensive literature review^{5,6}. In the transitional stage, both early (short-lived) and late (long-lived) species coexist, leading to high species diversity. In this stage, biomass is relatively low and resource level is still high, promoting higher productivity. In the late stage, stability may be high but diversity and productivity may be low because of competitive exclusions⁷ and sequestration of limited resources by accumulated litter and biomass^{5,12}.

The two communities compared at their site A represent two extreme degrees of disturbance: one undisturbed and the other heavily degraded. Although the undisturbed community did support relatively higher diversity and above-ground biomass than the highly degraded community, intermediate amounts of disturbance or transitional stages, which may well support even higher diversity and productivity, were not accounted for^{2–6} (Fig. 1).

Over space (that is, within one habitat), the 'intermediate disturbance' hypothesis⁷ predicts that diversity will be highest at an intermediate level of disturbance^{2,3}. Extensive evidence^{2–12} in support of this hypothesis is not in agreement with the conclusions of Bai *et al.*

Finally, over time at one locality, successional studies from various ecosystems — particularly those covering entire successional cycles — reveal that biodiversity and productivity are highest in the mid- or transitional-stage of succession, when both early- and late-stage species coexist. The high diversity and productivity then gradually decline owing to accumulated biomass and litter and therefore to increasing competition^{2–12} (Fig. 1).

Although each particular case would be expected to show some deviation from the general patterns in Fig. 1, because of the life history of dominant species, less destructive

disturbance, or variation in resources available over time¹³, for example, it is not clear why the Inner Mongolia grassland should be so different¹. If community stability, whose estimate depends on how it is measured, and biomass both increase with succession and are really high in undisturbed mature ecosystems, then the contrasting patterns between diversity and stability call their relationship into question.

The high stability in mature, or late-stage succession, grassland may be at least in part caused by the longer lifespan of the remaining, competitive perennial species (unlike annuals or short-lived plants in early succession) and by the high accumulated biomass, rather than by species diversity. For example, if community stability is measured as the coefficient of variation in biomass (CV, variance/mean)¹, then the CV, which is not independent of mean biomass¹⁴, will be lower when the biomass is higher, so the stability should be higher. Long-term, simultaneous monitoring of these variables in both above- and below-ground communities over the entire successional cycles of the grassland would help to clarify this point.

Qinfeng Guo

US Geological Survey, NPWRC, Jamestown, Nevada 89401, USA

e-mail: qguo@usgs.gov

1. Bai, Y., Han, X., Wu, J., Chen, Z. & Li, L. *Nature* **431**, 181–184 (2004).
2. Connell, J. H. *Science* **199**, 1302–1310 (1978).
3. Huston, M. *Biological Diversity* (Cambridge Univ. Press, Cambridge, 1994).
4. Gibson, D. J. & Hulbert, L. C. *Vegetatio* **72**, 175–185 (1987).
5. Guo, Q. *J. Veg. Sci.* **14**, 121–128 (2003).
6. *J. Veg. Sci.* http://www.opuluspress.se/pub/archives/JVS/archive_J.html
7. Grime, J. P. *Nature* **242**, 344–347 (1973).
8. Guo, Q. *Ambio* **32**, 428–430 (2003).
9. Lichter, J. *Ecol. Monogr.* **68**, 487–510 (1998).
10. Odum, E. P. *Science* **164**, 262–270 (1969).
11. Whittaker, R. H. *Communities and Ecosystems* 2nd edn (Macmillan, New York, 1975).
12. Ryan, M. G., Binkley, D., Fownes, J. H., Giardina, C. P. & Senock, R. S. *Ecol. Monogr.* **74**, 393–414 (2004).
13. Yang, L. H. *Science* **306**, 1565–1567 (2004).
14. Collins, S. L. *Trends Ecol. Evol.* **10**, 95–96 (1995).

doi:10.1038/nature03583

PLANT COMMUNITIES

Wu et al. reply

Reply to: Wang, S. *et al.* *Nature* doi:10.1038/nature03862 (2005) and Guo, Q. *Nature* doi:10.1038/nature03583 (2005)

Some of our findings¹ are questioned by Wang *et al.*², on the basis that we use inconsistent data and fail to distinguish spatial heterogeneity effects. Here we show that both claims are unfounded. We also address the questions raised by Guo³ concerning how the steppe communities vary as they mature.

In 1998 (not in 1999, as Wang *et al.* indicate), the sampling transect in our site A (the

Leymus chinensis community) was moved about 60 m northward within the same fenced plant community in a fairly uniform environmental setting to allow for more replicates and other sampling activities. To assure data consistency, we chose 5 quadrats in the new location at exactly the same distance interval as in the old transect. Our reanalysis of the 1980–97 data confirms our previous result¹ that live

above-ground biomass (LAB) was significantly correlated with January–July precipitation, but not with annual precipitation.

By contrast, Wang *et al.* show that LAB is correlated with April–August and annual precipitation². However, an earlier analysis⁴ by some of these authors using 1980–89 data showed that peak LAB correlated to annual precipitation only in the *Stipa grandis* community (our site B), not site A, contradicting their result presented here². Using the same 1980–89 data set, they reported that peak LAB of site A was significantly correlated with July precipitation and April–September precipitation⁴, but later found that above-ground net primary production was correlated with neither annual precipitation nor April–September precipitation⁵. Differences in timescale, biomass variables and data analysis procedures may have contributed to these seeming discrepancies, and the results of Wang *et al.* do not therefore invalidate our findings.

We are aware that spatial heterogeneity and scale influence ecological analysis^{6,7}, but the fine-grained within-community heterogeneity of site A is unlikely to have significantly undermined our results. The LAB and species composition estimated from 5 and 20 quadrats show quite similar results (for example, $\text{mean}_{\text{LAB},5\text{quadrats}} = 185.21$, $\text{s.d.}_{\text{LAB},5\text{quadrats}} = 30.97$, whereas $\text{mean}_{\text{LAB},20\text{quadrats}} = 188.05$, $\text{s.d.}_{\text{LAB},20\text{quadrats}} = 31.09$, on 30 August 2001). Also, our correlation analyses treated the 5 quadrats as replicates, and each data pair was always from the same quadrat over the 24 years. This was not ‘pooling’ all quadrats together, as suggested by Wang *et al.* Spatial variability precludes neither compensatory effects nor statistical analysis using replicates: on the contrary, statistical power should increase with the number of replicates.

Wang *et al.* criticize the use of relative biomass for detecting compensatory effects, but our correlation analyses were based on absolute biomass (although our Fig. 3 in ref. 1 used relative biomass for visual clarity). They suggest that the relatively small number of

species pairs showing negative correlations trivializes compensation effects, but this reasoning is flawed because the dynamics of even a few dominant species can significantly influence ecosystem productivity and its stability, as shown in our study¹.

Guo³ questions our finding that plant communities in the Inner Mongolia grassland (site A and site B) achieve high species richness, productivity and ecosystem stability concurrently at late successional stages¹. Although our focus was on the relationship between ecosystem stability and compensatory interactions in mature steppe communities, Guo’s comments raise questions concerning species diversity and biomass production over the course of succession.

Our conclusion did not refer to species richness and productivity as being the “highest”, nor did we suggest any particular relationship between richness and productivity along a succession gradient in space or time¹, so this cannot be in contradiction to the intermediate disturbance hypothesis. Existing studies in the same area indicate that species evenness (relative abundance), rather than richness (number of species), shows a hump-shaped response to a grazing gradient^{8,9}. It is therefore necessary to distinguish between species richness and species diversity, which usually incorporates both richness and evenness.

Guo argues that the highest biodiversity and productivity should occur before plant communities reach their mature state, a point articulated several decades ago^{6,10,11}. Relevant studies of the Inner Mongolia grassland^{12,13} show that a hump-shaped relationship of species richness is not evident with respect to either succession stages or community biomass. We acknowledge that such relationships may depend on the timescale of observation. Could the species richness and biomass of our two study sites decline several decades from now if they continue to be enclosed against fires and grazing by large animals? They might, but this is only speculation.

We disagree with Guo’s contention that the

coefficient of variation (CV, variance/mean) in community biomass decreases with increasing mean biomass. Usually when mean biomass increases, so does the variance. Thus, CV does not necessarily decrease and, in general, the main reason for using CV instead of variance itself is to minimize bias introduced by changing the mean.

We concur that the longer lifespan of perennial species may contribute to the higher stability of mature steppes. But we note that both undisturbed mature and degraded successional communities in the Inner Mongolia grassland contain annuals as well as perennials^{12,13}, except in extreme situations where top soils are largely lost. We also agree that monitoring both above- and below-ground ecological variables is important for fully understanding the relationship between biodiversity and ecosystem processes.

Jianguo Wu*†, Yongfei Bai*, Xingguo Han*,
Linghao Li*, Zuozhong Chen*

*Laboratory of Quantitative Vegetation Ecology,
Institute of Botany, Chinese Academy of Sciences,
Beijing 100093, China

†Faculty of Ecology, Evolution and Environmental
Science, School of Life Sciences, Arizona State
University, Tempe, Arizona 85287-4501, USA
e-mail: jingle.wu@asu.edu

1. Bai, Y., Han, X., Wu, J., Chen, Z. & Li, L. *Nature* **431**, 181–184 (2004).
2. Wang, S. *et al. Nature* doi:10.1038/nature03862 (2005).
3. Guo, Q. *Nature* doi:10.1038/nature03583 (2005).
4. Xiao, X. M., Wang, Y., Jiang, S., Ojima, D. S. & Bonham, C. D. *J. Arid Environ.* **31**, 283–299 (1995).
5. Xiao, X. M., Jiang, S., Wang, Y., Ojima, D. S. & Bonham, C. D. *Vegetatio* **123**, 1–12 (1996).
6. Wu, J. G. & Loucks, O. L. *Q. Rev. Biol.* **70**, 439–466 (1995).
7. Wu, J. G. *Landscape Ecol.* **19**, 125–138 (2004).
8. Li, Y. H. *Acta Phytocool. Geobot. Sinica* **12**, 189–196 (1988).
9. Li, Y. H. *Acta Botanica Sinica* **35**, 877–884 (1993).
10. Odum, E. P. *Science* **164**, 262–270 (1969).
11. Loucks, O. L. *Am. Zool.* **10**, 17–25 (1970).
12. Wang, W., Liu, Z. L., Hao, D. Y. & Liang, C. Z. *Acta Phytocool. Sinica* **20**, 449–459 (1996).
13. Wang, W., Liu, Z. L., Hao, D. Y. & Liang, C. Z. *Acta Phytocool. Sinica* **20**, 460–471 (1996).

doi:10.1038/nature03584

Vortices and superfluidity in a strongly interacting Fermi gas

M. W. Zwierlein¹, J. R. Abo-Shaeer^{1†}, A. Schirotzek¹, C. H. Schunck¹ & W. Ketterle¹

Quantum degenerate Fermi gases provide a remarkable opportunity to study strongly interacting fermions. In contrast to other Fermi systems, such as superconductors, neutron stars or the quark-gluon plasma of the early Universe, these gases have low densities and their interactions can be precisely controlled over an enormous range. Previous experiments with Fermi gases have revealed condensation of fermion pairs. Although these and other studies were consistent with predictions assuming superfluidity, proof of superfluid behaviour has been elusive. Here we report observations of vortex lattices in a strongly interacting, rotating Fermi gas that provide definitive evidence for superfluidity. The interaction and therefore the pairing strength between two ^6Li fermions near a Feshbach resonance can be controlled by an external magnetic field. This allows us to explore the crossover from a Bose-Einstein condensate of molecules to a Bardeen-Cooper-Schrieffer superfluid of loosely bound pairs. The crossover is associated with a new form of superfluidity that may provide insights into high-transition-temperature superconductors.

The first observations of Bose-Einstein condensates (BECs) of molecules consisting of loosely bound fermionic atoms^{1–3} initiated a series of explorations^{4–12} of the crossover between a BEC and a Bardeen-Cooper-Schrieffer (BCS) superfluid^{13–15}. When an external magnetic field is varied across a Feshbach resonance, these molecules transform adiabatically into the Cooper pairs of a BCS superfluid. All physical properties are expected to vary smoothly throughout this crossover. For example, the size of fermion pair condensates smoothly increases from the BEC- to the BCS-side of the resonance, while the strength of the bond between two paired atoms smoothly decreases. The similar size and shape of normal and condensed gas clouds makes it difficult to detect condensation on the BCS-side. However, using a rapid magnetic field sweep back to the BEC-side, introduced by the Boulder group⁴, pair condensation was observed^{4,5,12}. Although Bose-Einstein condensation and superfluidity are intimately connected, they do not necessarily occur together. In lower dimensions, superfluidity occurs in the absence of Bose-Einstein condensation¹⁶. An ideal Bose gas or disordered three-dimensional system can have a condensate, but shows no superfluidity¹⁶. Phase fluctuations, which are probably present in short-lived condensates of fermionic pairs¹, can suppress superfluid behaviour.

Several ground-breaking studies in Fermi gases of hydrodynamic expansion^{7,17}, collective excitations^{8,9}, thermodynamic properties¹¹ and the binding energy of pairs¹⁰ were suggestive of superfluid behaviour or were consistent with theoretical calculations predicting superfluidity, but did not provide unambiguous evidence. In the meantime, several theoretical papers^{18–22} emphasized that the rotational properties of a gas of fermion pairs could directly reveal superfluidity in such systems.

Quantized vortices in a rotating gas provide conclusive evidence for superfluidity because they are a direct consequence of the existence of a macroscopic wavefunction that describes the superfluid. The velocity field of the superfluid is proportional to the gradient of the wavefunction's phase. In such a case, flow must be irrotational and angular momentum can enter the system only in the

form of discrete line defects (vortices). In contrast, for a normal gas the lowest state of rotation corresponds to rigid body rotation. Metastable vortex patterns have been observed in classical inviscid fluids²³. However, the final number and charge of the vortices depends chaotically on the initial conditions, in contrast to the regular vortex lattices that we have reproducibly observed. Furthermore, vortex patterns in classical fluids are only stable at extremely low viscosity (that is, for Reynolds numbers $>10^5$). For a Boltzmann gas rotating close to the trap frequency, the Reynolds number is approximately the cloud size divided by the mean free path, which does not exceed 10^3 in our case. Pauli blocking can only decrease the Reynolds number further.

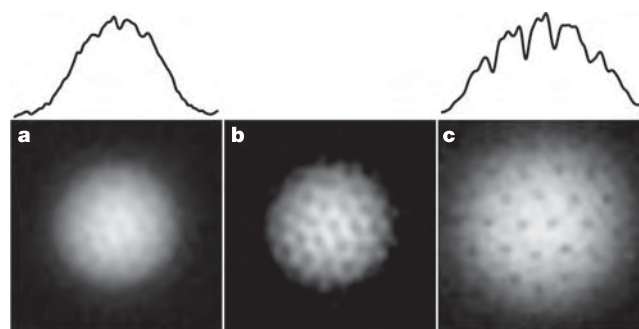


Figure 1 | Observation of a vortex lattice in a molecular condensate. **a**, Fixed field. Stirring for 800 ms, followed by 400 ms of equilibration, and imaging after 12 ms time-of-flight all took place at 766 G. The vortex core depletion of the integrated density profile is barely 10%, as indicated by the 5- μm -wide cut on top. **b**, Fourier filter applied to **a** to accentuate the vortex contrast. Spatial frequencies with an absolute value of about the inverse vortex core size were enhanced by a factor of four. **c**, Varying field. The vortex lattice was created at 766 G and imaged at 735 G following the procedure outlined in the text. The vortex core depletion is now about 35% (see 5- μm -wide cut on top). The field of view is 780 $\mu\text{m} \times 780 \mu\text{m}$.

¹Department of Physics, MIT-Harvard Center for Ultracold Atoms, and Research Laboratory of Electronics, MIT, Cambridge, Massachusetts 02139, USA. [†]Present address: Lawrence Berkeley National Laboratory, One Cyclotron Road, MS 88R0192, Berkeley, California 94720, USA.

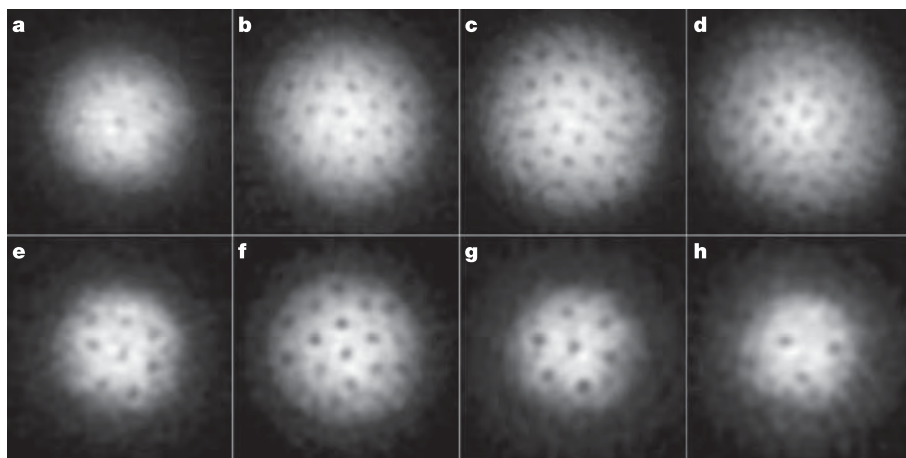


Figure 2 | Vortices in a strongly interacting gas of fermionic atoms on the BEC- and the BCS-side of the Feshbach resonance. At the given field, the cloud of lithium atoms was stirred for 300 ms (a) or 500 ms (b–h) followed by an equilibration time of 500 ms. After 2 ms of ballistic expansion, the

magnetic field was ramped to 735 G for imaging (see text for details). The magnetic fields were 740 G (a), 766 G (b), 792 G (c), 812 G (d), 833 G (e), 843 G (f), 853 G (g) and 863 G (h). The field of view of each image is $880 \mu\text{m} \times 880 \mu\text{m}$.

Experimental procedure

To create a strongly interacting Fermi gas, spin-polarized fermionic ^6Li atoms were sympathetically cooled to degeneracy by ^{23}Na atoms in a magnetic trap²⁴. The Fermi cloud was then loaded into an optical dipole trap, and an 875 G external magnetic field was applied. Here a 50%/50% spin mixture of the two lowest hyperfine states of ^6Li was prepared. Between these two states, labelled $|1\rangle$ and $|2\rangle$, there is a 300-G-wide Feshbach resonance located at 834 G (refs 25, 26). Evaporative cooling (achieved by reducing the laser power) accompanied by a magnetic field ramp to 766 G on the BEC-side of the resonance typically produced a BEC of 3×10^6 molecules³.

Previous experiments studying the rotation of atomic BECs employed magnetic traps operating at low bias fields^{27–31}. Because the Feshbach resonance in our system occurs between two high-field seeking states that cannot be trapped magnetically, an optical dipole trap operating at high magnetic bias fields was necessary. Our set-up employed a trapping beam with a $1/e^2$ radius of $123 \mu\text{m}$ (wavelength $1,064 \text{ nm}$), radially confining the gas with a trap frequency of 59 Hz at a power of 145 mW . Axial confinement with trap frequency $\nu_z = 23 \text{ Hz}$ was provided by an applied magnetic field curvature that decreased the radial trap frequency to $\nu_r = 57 \text{ Hz}$. The aspect ratio of the trap was 2.5. In this trap, at a field of 766 G , condensates of 1×10^6 molecules (the typical number in our experiment after rotating the cloud) have Thomas–Fermi radii of about $45 \mu\text{m}$ radially and $110 \mu\text{m}$ axially, a peak molecular density of $2.6 \times 10^{12} \text{ cm}^{-3}$, a chemical potential of about 200 nK , and a characteristic microscopic length scale of $1/k_F \approx 0.3 \mu\text{m}$. Here, the Fermi wavevector k_F is defined by the Fermi energy (E_F) of a non-interacting two-state mixture of ^6Li atoms of mass m with total atom number N in a harmonic trap of (geometric) mean frequency $\bar{\omega}$, $E_F = \hbar\bar{\omega}(3N)^{1/3} \equiv \hbar^2 k_F^2/2m$. Throughout this Article we will estimate the interaction parameter $1/k_F a$ using the average number of fermion pairs $N/2 = 1 \times 10^6$. Here, a is the scattering length between atoms in states $|1\rangle$ and $|2\rangle$. At a field of 766 G , $1/k_F a = 1.3$. Because this gas is strongly interacting, it is difficult to extract a temperature from the spatial profile. For weaker interactions (at 735 G) the condensate fraction was in excess of 80%, which would isentropically connect to an ideal Fermi gas³² at $T/T_F = 0.07$. The BEC–BCS crossover ($1/k_F |a| < 1$) occurs in the region between 780 G and 925 G .

The trapped cloud was rotated about its long axis using a blue-detuned laser beam (wavelength 532 nm)^{28,29,33}. A two-axis acousto-optic deflector generated a two-beam pattern (beam separation $d = 60 \mu\text{m}$, gaussian beam waist $w = 16 \mu\text{m}$) that was rotated symmetrically around the cloud at a variable angular frequency Ω .

The two beams with 0.4 mW power each produced a repulsive potential of 125 nK for the ^6Li cloud, creating a strongly anisotropic potential. This method was first tested using a weakly interacting, atomic BEC of ^{23}Na in the stretched upper hyperfine state in an optical trap with $\nu_r = 60 \text{ Hz}$, $\nu_z = 23 \text{ Hz}$. Fully equilibrated lattices of up to 80 vortices were observed. The vortex number decayed with a $1/e$ lifetime of $4.2 \pm 0.2 \text{ s}$, while the atom number decayed, owing to three-body losses and evaporation, with a lifetime of $8.8 \pm 0.4 \text{ s}$. The roundness of the optical trap and its alignment with both the optical stirrer and the axes of the magnetic potential were critical. Any deviation from cylindrical symmetry owing to misalignment, optical aberrations, or gravity rapidly damped the rotation. The generation of vortices in sodium was comparatively forgiving, and had to be optimized before vortices in $^6\text{Li}_2$ could be observed.

Observation of vortex lattices

In experiments with ^6Li close to the Feshbach resonance, the interaction strength between atoms in states $|1\rangle$ and $|2\rangle$ can be freely tuned via the magnetic field. Thus, it is possible to choose different magnetic fields to optimize the three steps involved in the creation of a vortex lattice: stirring of the cloud (for 800 ms at a typical stirring frequency of 45 Hz), the subsequent equilibration (typically 500 ms) and time-of-flight expansion for imaging. To stay close to the

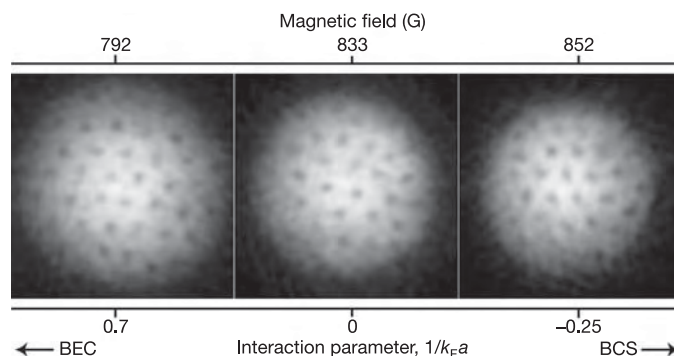


Figure 3 | Optimized vortex lattices in the BEC–BCS crossover. After a vortex lattice was created at 812 G , the field was ramped in 100 ms to 792 G (BEC-side), 833 G (resonance) and 853 G (BCS-side), where the cloud was held for 50 ms . After 2 ms of ballistic expansion, the magnetic field was ramped to 735 G for imaging (see text for details). The field of view of each image is $880 \mu\text{m} \times 880 \mu\text{m}$.

analogous case of an atomic condensate, our search for vortices started on the BEC-side of the resonance, at a fixed magnetic field of 766 G. To image the cloud, the trapping beam was switched off and the cloud expanded in the residual magnetic potential. After 12 ms time-of-flight, a molecular absorption image along the axial direction was taken with light resonantly exciting atoms in state $|2\rangle^3$ (Fig. 1a). Although the contrast of the vortex cores was low, a regular lattice pattern containing about 25 vortices was visible (see Fourier filtered image in Fig. 1b). This establishes superfluidity for molecular condensates.

Subsequently, it was found that the contrast of the vortex cores could be enhanced by the following procedure: the magnetic field curvature was reduced by a factor of five during the first millisecond of time-of-flight. After 2 ms expansion at the initial field, the magnetic field was ramped down over 2 ms to 735 G. The cloud was imaged after an additional 9 ms of expansion at this field. Attempts to image at even lower fields did not enhance the contrast. Further improvements were achieved by forcing the cloud to expand faster by increasing the power of the optical trap by a factor of 4.5 during the last 2 ms of trapping (Fig. 1c). We suspect that owing to the residual magnetic field curvature, a faster expansion was superior to a longer time-of-flight.

Using this procedure, we observed vortices that were created not only on the BEC- but also the BCS-side of the Feshbach resonance, at magnetic fields between 740 G and 863 G (Fig. 2). On the BCS-side, isolated fermion pairs are unstable. As the cloud expands from the trap and the density decreases, the pairs will become more fragile and dissociate at a certain point in the time-of-flight. Information on the centre-of-mass wavefunction of the pairs, and hence the vortex contrast, will be gradually lost. The field ramp to the BEC-side during expansion protects the pairs by transforming them into stable molecules. At 853 G, this ramp could be delayed by 6 ms into ballistic expansion before the vortex contrast was lost. It is not clear why this ramp was found to be necessary already at 812 G, on the BEC-side of the resonance, where isolated fermion pairs are stable.

It is impossible for vortex lattices to form during ballistic expansion at the imaging field (735 G). We show below that the formation of vortex lattices even at the high density of the trapped cloud takes several hundred milliseconds. Furthermore, even if there was some

unpredicted fast formation mechanism for vortices, they could not form a regular array with long-range order in a cloud that expands at the speed of sound of the trapped gas. The observation of vortex lattices on the BCS-side of the Feshbach resonance above 834 G demonstrates superfluidity of fermions at magnetic fields where they cannot form stable, isolated molecules.

The highest number of vortices (~ 40) was obtained by stirring at 766 G and then equilibrating close to resonance at 812 G ($1/k_F a = 0.35$). We suspect that the violent nature of the stirring produced more heating near the Feshbach resonance where the pairs are loosely bound. On the other hand, fields closer to resonance were favourable for equilibration owing to suppression of vibrational relaxation. After preparing such a vortex lattice at 812 G, the magnetic field was ramped over 100 ms to a test field. After a hold time of 50 ms, the vortex lattice was imaged as discussed above (Fig. 3). Vortices were observed for test fields between 700 G ($1/k_F a = 3.8$) and 954 G ($1/k_F a = -1.2$) (Fig. 4).

The regularity of the lattice proves that all vortices have the same vorticity. From their number, the size of the cloud and the quantum of circulation $h/2m$ for each vortex, we can estimate the rotational frequency of the lattice. For an optimized stirring procedure, we find that it is close to the stirring frequency. This excludes a quantum of circulation of h/m or doubly charged vortices.

Formation and lifetime of vortex lattices

Before we found the detection scheme described above, we studied the formation and decay of the vortex lattices using a different procedure. The magnetic field was lowered to 735 G ($1/k_F a = 2.3$) in the last 30 ms before expansion. Reduction of the magnetic field curvature by a factor of five took place during the last 5 ms before expansion, avoiding undesired compression of the cloud in the axial direction. As before, the trap was compressed by increasing the trapping power by a factor of 4.5 during the last 2 ms before the switch-off. Imaging was done after 12 ms expansion at 735 G.

When a superfluid is rotated, the creation of quantized vortices is energetically favoured only above a certain critical rotation frequency. In some cases, a higher rotational frequency is necessary to actually nucleate the vortices³⁴. This occurs through a dynamic instability of surface excitations^{33,35,36}. Figure 5 shows that vortices in the BEC region were created over a large range of stirring frequencies, as opposed to only near the quadrupole surface mode resonance^{28,29}. Similar non-resonant behaviour was observed in atomic sodium condensates with small stirring beams³⁶. The strong dependence of the observed stirring efficiency curve on the magnetic

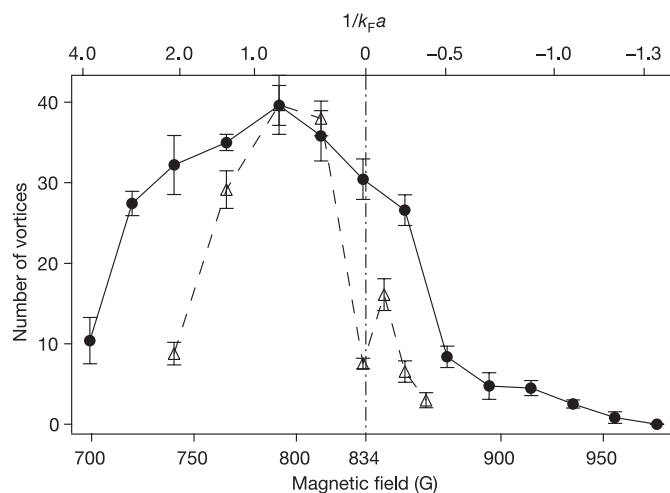


Figure 4 | Vortex number versus magnetic field and interaction strength in the BEC-BCS crossover. The open triangles show the number of vortices obtained after stirring and equilibration at the given, fixed magnetic field, as in Fig. 2. For the filled circles, a procedure similar to the one in Fig. 3 was used, where the vortex lattice was prepared at 812 G, and then the field was ramped to the test field. The position of the Feshbach resonance²⁶ is marked with the vertical dash-dotted line. The data points and error bars give the average and standard deviation of several measurements.

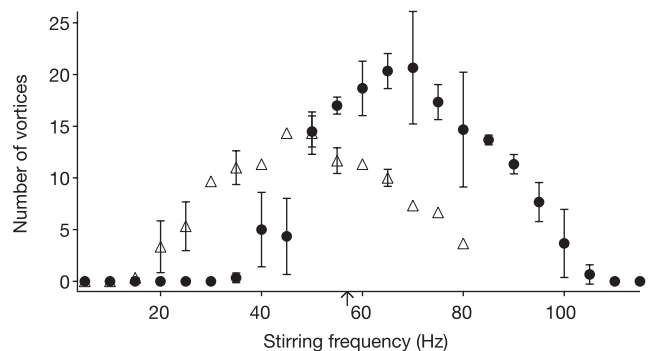


Figure 5 | Vortex number versus stirring frequency in the BEC region for different interaction strengths. Vortices were efficiently created over a broad range of stirring frequencies. The open triangles and filled circles correspond to stirring (for 800 ms) and equilibration (for 500 ms) at 766 G and 812 G, respectively. The data points and error bars (if larger than the symbol) show the average and standard deviation of three measurements. The radial trap frequency (57 Hz) is indicated by the small arrow on the horizontal axis.

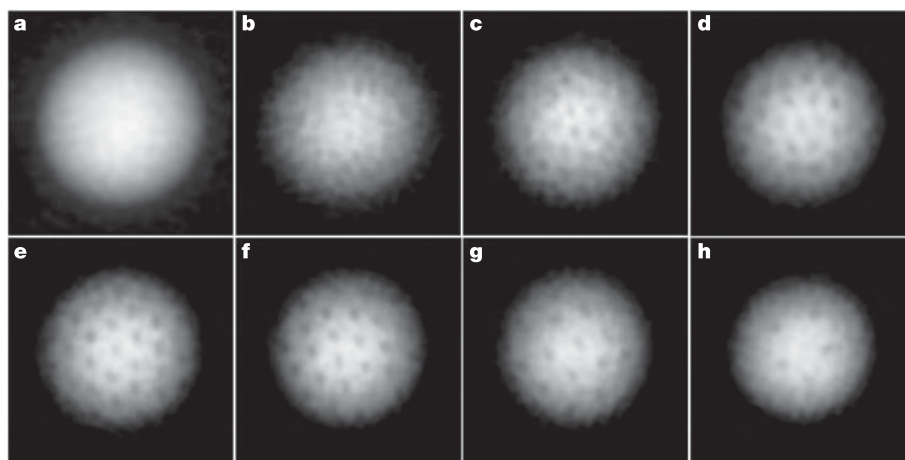


Figure 6 | Formation and decay of a vortex lattice in a fermion pair condensate on the BEC-side close to the Feshbach resonance. A molecular condensate, prepared at 766 G as shown in **a**, was stirred for 800 ms. The field was then ramped to 812 G in 20 ms for equilibration. At this field, $1/k_F a = 0.35$, and the condensate was deep in the strongly interacting regime. To observe the vortex lattice, the field was ramped in 25 ms to 735 G

($1/k_F a = 2.3$), where the condensate was released from the trap and imaged after 12 ms time-of-flight. The equilibration times after the end of the stirring were 40 ms (**b**), 240 ms (**c**), 390 ms (**d**), 790 ms (**e**), 1,140 ms (**f**), 1,240 ms (**g**) and 2,940 ms (**h**). Owing to stirring, evaporation and vibrational relaxation, the number of fermion pairs decayed from 3×10^6 (**a**) to 1×10^6 (**b–h**). The field of view of each image is $830 \mu\text{m} \times 830 \mu\text{m}$.

field strength could be related to the increase of the speed of sound in the condensate for stronger interactions, leading to a higher critical velocity.

Figure 6 shows the formation of a vortex lattice close to the Feshbach resonance, on the BEC-side at 812 G ($1/k_F a = 0.35$). Immediately after stirring, the cloud was in a turbulent state (Fig. 6b). It took several hundred milliseconds for the vortices to fully crystallize into a lattice (Fig. 6c–e). The vortices arranged themselves in a hexagonal Abrikosov lattice to minimize their interaction energy^{29,37}. Because the trap potential was not perfectly round (trap asymmetry $(v_x^2 - v_y^2)/(v_x^2 + v_y^2) \approx 0.03$), the vortex lattice slowly decayed on a timescale of several seconds (Fig. 6f–h). These observations are fully analogous to those already made in atomic BECs^{29,35,38}. The formation time of several hundred milliseconds is in agreement with these experimental studies, as well as with a recent theoretical study on the vortex lattice formation in a strongly interacting Fermi gas³⁹. Note that this timescale was found to be independent of temperature³⁸, and seems to represent an intrinsic timescale of superfluid hydrodynamics. The long formation time excludes the possibility that ordered vortex lattices can be created during the 30 ms field ramp to 735 G, used for imaging. This holds even more strongly for the imaging method described first, where the magnetic field was switched only in time-of-flight. We are not aware of any possible formation mechanism that could create regular vortex lattices during ballistic expansion.

In principle, the presence of vortices can be used to map out the superfluid regime as a function of temperature and interaction strength, even in regions where fermion pairs are much larger than the interparticle spacing and can no longer be detected by transforming them into stable molecules^{4,5,12}. As a first step, the lifetime of the vortex lattice was studied at different magnetic fields and interaction strengths. After preparing a fully crystallized vortex lattice containing about 30 vortices in the BEC-region at 812 G, the magnetic field was ramped to a chosen point in the crossover. After a variable hold time, the cloud was imaged at 735 G as described above, and the remaining number of vortices was counted. The results of this measurement are summarized in Fig. 7. The longest lifetime, 3.4 s, was obtained for magnetic fields slightly below resonance, near 810 G ($1/k_F a = 0.4$). Here, we observed four vortices even after 7 s. During this time the fluid at the vortex core (with characteristic size $1/k_F$) rotated more than 50,000 times, displaying truly superfluid behaviour. As expected, the lifetime was reduced at low magnetic fields where the

molecules heat up owing to vibrational relaxation. In addition, and unexpectedly, a narrow dip in lifetime at approximately 831 G with a width of 8 G was observed. We speculate that this decrease in lifetime so close to resonance is caused by a coupling of the external motion of the loosely bound pairs to their internal motion, causing pair breaking or rotational excitation. Indeed, at 831 G, the binding energy for isolated molecules divided by $\hbar$ is about 35 Hz, comparable to the initial rotational frequency of the lattice. The reduced lifetime on the BCS-side could be caused by an increasing fraction of thermal fermion pairs^{5,38} due to the decreasing critical temperature for superfluidity. In a two-fluid model, angular momentum is stored as a vortex lattice in the superfluid component. Friction is provided by the normal component, which increases as we move further away from resonance on the BCS-side. At 925 G ($1/k_F a = -1.0$) the vortex lifetime was reduced to only 10 ms, with a large error bar indicating

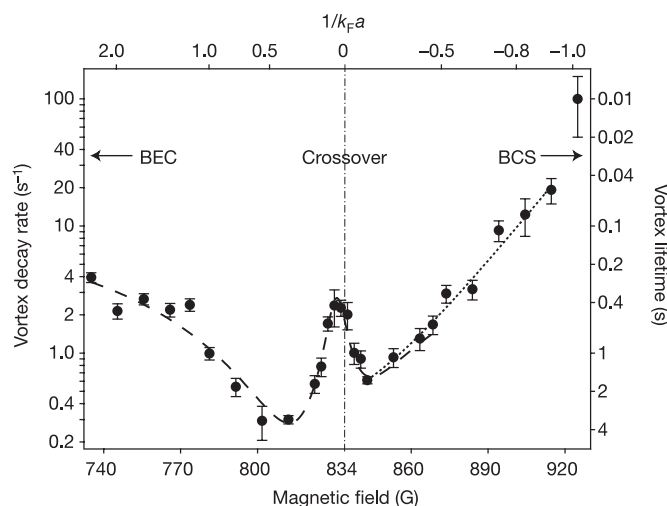


Figure 7 | Decay rate and lifetime of the vortex lattice versus magnetic field and interaction strength. The data points and error bars give the decay rate or lifetime and one standard deviation extracted from exponential fits to the vortex number decay. The dashed and dotted lines are gaussian and exponential fits to guide the eye, the dashed line including a lorentzian fit for the narrow feature near resonance. The position of the Feshbach resonance²⁶ is marked with the vertical dash-dotted line.

strong fluctuations in the vortex number. Here we might be approaching the region where the superfluid-to-normal transition takes place.

Conclusions

We have detected long-lived vortex lattices in a strongly interacting Fermi gas over the entire BEC–BCS crossover region by imaging them after switching to lower magnetic fields during ballistic expansion. This provides the first direct signature of superfluidity in these systems. We expect that vortices in rotating Fermi gases will serve as an important starting point for future studies on superfluid dynamics.

Received 9 May; accepted 31 May 2005.

- Greiner, M., Regal, C. A. & Jin, D. S. Emergence of a molecular Bose–Einstein condensate from a Fermi gas. *Nature* **426**, 537–540 (2003).
- Jochim, S. *et al.* Bose–Einstein condensation of molecules. *Science* **302**, 2101–2103 (2003).
- Zwierlein, M. W. *et al.* Observation of Bose–Einstein condensation of molecules. *Phys. Rev. Lett.* **91**, 250401 (2003).
- Regal, C. A., Greiner, M. & Jin, D. S. Observation of resonance condensation of fermionic atom pairs. *Phys. Rev. Lett.* **92**, 040403 (2004).
- Zwierlein, M. W. *et al.* Condensation of pairs of fermionic atoms near a Feshbach resonance. *Phys. Rev. Lett.* **92**, 120403 (2004).
- Bartenstein, M. *et al.* Crossover from a molecular Bose–Einstein condensate to a degenerate Fermi gas. *Phys. Rev. Lett.* **92**, 120401 (2004).
- Bourdel, T. *et al.* Experimental study of the BEC–BCS crossover region in lithium 6. *Phys. Rev. Lett.* **93**, 050401 (2004).
- Kinast, J., Hemmer, S. L., Gehm, M. E., Turlapov, A. & Thomas, J. E. Evidence for superfluidity in a resonantly interacting Fermi gas. *Phys. Rev. Lett.* **92**, 150402 (2004).
- Bartenstein, M. *et al.* Collective excitations of a degenerate gas at the BEC–BCS crossover. *Phys. Rev. Lett.* **92**, 203201 (2004).
- Chin, C. *et al.* Observation of the pairing gap in a strongly interacting Fermi gas. *Science* **305**, 1128–1130 (2004).
- Kinast, J. *et al.* Heat capacity of a strongly-interacting Fermi gas. *Science* **307**, 1296–1299 (2005).
- Zwierlein, M. W., Schunck, C. H., Stan, C. A., Raupach, S. M. F. & Ketterle, W. Formation time of a fermion pair condensate. *Phys. Rev. Lett.* **94**, 180401 (2005).
- Eagles, D. M. Possible pairing without superconductivity at low carrier concentrations in bulk and thin-film superconducting semiconductors. *Phys. Rev.* **186**, 456–463 (1969).
- Leggett, A. J. in *Modern Trends in the Theory of Condensed Matter* (eds Pekalski, A. & Przystawa, J.) 13–27 (Proc. XVIth Karpacz Winter School of Theoretical Physics, Springer, Berlin, 1980).
- Nozières, P. & Schmitt-Rink, S. Bose condensation in an attractive fermion gas: from weak to strong coupling superconductivity. *J. Low Temp. Phys.* **59**, 195–211 (1985).
- Huang, K. in *Bose–Einstein Condensation* (eds Griffin, A., Snoke, D. W. & Stringari, S.) 31–50 (Cambridge Univ. Press, Cambridge, 1995).
- O'Hara, K. M., Hemmer, S. L., Gehm, M. E., Granade, S. R. & Thomas, J. E. Observation of a strongly interacting degenerate Fermi gas of atoms. *Science* **298**, 2179–2182 (2002).
- Rodriguez, M., Paraoanu, G.-S. & Törmä, P. Vortices in trapped superfluid Fermi gases. *Phys. Rev. Lett.* **87**, 100402 (2001).
- Bruun, G. M. & Viverit, L. Vortex state in superfluid trapped Fermi gases at zero temperature. *Phys. Rev. A* **64**, 063606 (2001).
- Pitaevskii, L. & Stringari, S. The quest for superfluidity in Fermi gases. *Science* **298**, 2144–2146 (2002).
- Cozzini, M. & Stringari, S. Fermi gases in slowly rotating traps: superfluid versus collisional hydrodynamics. *Phys. Rev. Lett.* **91**, 070401 (2002).
- Bulgac, A. & Yu, Y. Vortex state in a strongly coupled dilute atomic fermionic superfluid. *Phys. Rev. Lett.* **91**, 190404 (2003).
- Schecter, D. A., Dubin, D. H. E., Fine, K. S. & Driscoll, C. F. Vortex crystals from 2D Euler flow: Experiment and simulation. *Phys. Fluids* **11**, 905–914 (1999).
- Hadzibabic, Z. *et al.* Fifty-fold improvement in the number of quantum degenerate fermionic atoms. *Phys. Rev. Lett.* **91**, 160401 (2003).
- Dieckmann, K. *et al.* Decay of ultracold fermionic lithium gas near a Feshbach resonance. *Phys. Rev. Lett.* **89**, 203201 (2002).
- Bartenstein, M. *et al.* Precise determination of ^6Li cold collision parameters by radio-frequency spectroscopy on weakly bound molecules. *Phys. Rev. Lett.* **94**, 103201 (2004).
- Matthews, M. R. *et al.* Vortices in a Bose–Einstein condensate. *Phys. Rev. Lett.* **83**, 2498–2501 (1999).
- Madison, K. W., Chevy, F., Wohlleben, W. & Dalibard, J. Vortex formation in a stirred Bose–Einstein condensate. *Phys. Rev. Lett.* **84**, 806–809 (2000).
- Abo-Shaeer, J. R., Raman, C., Vogels, J. M. & Ketterle, W. Observation of vortex lattices in Bose–Einstein condensates. *Science* **292**, 476–479 (2001).
- Haljan, P. C., Coddington, I., Engels, P. & Cornell, E. A. Driving Bose–Einstein-condensate vorticity with a rotating normal cloud. *Phys. Rev. Lett.* **87**, 210403 (2001).
- Hodby, E., Hechenblaikner, G., Hopkins, S. A., Maragò, O. M. & Foot, C. J. Vortex nucleation in Bose–Einstein condensates in an oblate, purely magnetic potential. *Phys. Rev. Lett.* **88**, 010405 (2002).
- Carr, L. D., Shlyapnikov, G. V. & Castin, Y. Achieving a BCS transition in an atomic Fermi gas. *Phys. Rev. Lett.* **92**, 150404 (2004).
- Onofrio, R. *et al.* Surface excitations in a Bose–Einstein condensate. *Phys. Rev. Lett.* **84**, 810–813 (2000).
- Anglin, J. R. Local vortex generation and the surface mode spectrum of large Bose–Einstein condensates. *Phys. Rev. Lett.* **87**, 240401 (2001).
- Madison, K. W., Chevy, F., Bretin, V. & Dalibard, J. Stationary states of a rotating Bose–Einstein condensate: routes to vortex nucleation. *Phys. Rev. Lett.* **86**, 4443–4446 (2001).
- Raman, C., Abo-Shaeer, J. R., Vogels, J. M., Xu, K. & Ketterle, W. Vortex nucleation in a stirred Bose–Einstein condensate. *Phys. Rev. Lett.* **87**, 210402 (2001).
- Madison, K. W., Chevy, F., Wohlleben, W. & Dalibard, J. Vortices in a stirred Bose–Einstein condensate. *J. Mod. Opt.* **47**, 2715–2723 (2000).
- Abo-Shaeer, J. R., Raman, C. & Ketterle, W. Formation and decay of vortex lattices in Bose–Einstein condensates at finite temperatures. *Phys. Rev. Lett.* **88**, 070409 (2002).
- Tonini, G. & Castin, Y. Formation of a vortex lattice in a rotating BCS Fermi gas. Preprint at (<http://arxiv.org/cond-mat/0504612>) (2005).

Acknowledgements We thank P. Zarth for experimental assistance and C. Stan for contributions in the early stages of the experiment. We also acknowledge discussions with the participants of the OCTS conference in Ohio, and thank J. Anglin, Z. Hadzibabic, D. Kleppner and A. Leanhardt for a critical reading of the manuscript. This work was supported by the NSF, ONR, ARO and NASA.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.W.Z. (zwierlei@mit.edu).

ARTICLES

Stargazin modulates AMPA receptor gating and trafficking by distinct domains

Susumu Tomita¹, Hillel Adesnik^{2*}, Masayuki Sekiguchi^{4*}, Wei Zhang^{3*}, Keiji Wada⁴, James R. Howe³, Roger A. Nicoll^{1,2} & David S. Bredt¹

AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptors mediate fast excitatory synaptic transmission in the brain. These ion channels rapidly deactivate and desensitize, which determine the time course of synaptic transmission. Here, we find that the AMPA receptor interacting protein, stargazin, not only mediates AMPA receptor trafficking but also shapes synaptic responses by slowing channel deactivation and desensitization. The cytoplasmic tail of stargazin determines receptor trafficking, whereas the ectodomain controls channel properties. Stargazin alters AMPA receptor kinetics by increasing the rate of channel opening. Disrupting the interaction of stargazin ectodomain with hippocampal AMPA receptors alters the amplitude and shape of synaptic responses, establishing a crucial function for stargazin in controlling the efficacy of synaptic transmission in the brain.

Most excitatory synapses in brain use glutamate as a neurotransmitter. The three major classes of glutamate-gated ion channel receptors—AMPA, kainate and NMDA (*N*-methyl-D-aspartate) types—subserve distinct functions^{1–5}. The AMPA-preferring glutamate receptors (AMPARs), which are also sensitive to kainate, are monovalent cation channels and mediate most of the post-synaptic depolarization that induces neuronal firing. High-affinity kainate-preferring receptors also participate in glutamate-mediated transmission at a subset of synapses. NMDA receptors (NMDARs) are permeable to calcium as well as monovalent cations. Calcium influx through NMDARs induces synaptic plasticity by changing the number of synaptic AMPARs. Because alterations in synaptic AMPAR trafficking underlie synaptic plasticity and provide a mechanism for information storage in the brain, intensive studies have addressed protein interactions with AMPARs^{6–10}.

The four-pass transmembrane protein stargazin traffics AMPARs^{11,12}. Stargazin is mutated in stargazer mice¹³, which show absence epilepsy and a lack of functional AMPARs in cerebellar granule cells^{11,12,14}. Although AMPAR function is normal in the forebrain of stargazer mice, three related transmembrane AMPAR regulatory proteins (TARPs) expressed in forebrain can also traffic AMPARs¹⁵. So far, the effects of stargazin on AMPAR function have been interpreted in terms of trafficking alone. We report here that stargazin functions in a similar way to auxiliary subunits of voltage-dependent calcium channels^{16,17} and controls both receptor trafficking and channel properties. By controlling AMPAR opening, stargazin shapes the postsynaptic currents that determine excitatory transmission.

Stargazin increases AMPAR trafficking and currents

Stargazin and related TARPs greatly increase glutamate-evoked currents from *Xenopus laevis* oocytes injected with limiting amounts of glutamate receptor 1 (GluR1)^{18,19}. To quantify further this effect, we measured glutamate-evoked currents in oocytes injected with varying amounts of stargazin and GluR1 (Fig. 1a, b). We found that

GluR1 levels that exhibit no current on their own evoke near-maximal current in the presence of stargazin (Fig. 1a, b). To determine whether stargazin influences both receptor trafficking and channel properties, we developed an assay to measure these effects independently. To study trafficking, we engineered a haemagglutinin (HA) epitope into the extracellular region of GluR1-flip and quantified surface expression by chemiluminescence²⁰. We detected decreasing GluR1 surface expression and glutamate-evoked currents in oocytes injected with 2, 1 or 0.1 ng GluR1 complementary RNA (cRNA). Co-injection of stargazin cRNA with 0.1 ng GluR1 greatly potentiated glutamate-evoked currents (Fig. 2a). Notably, the stargazin-mediated increase in glutamate-evoked currents was much greater than was the increase in GluR1 surface expression (Fig. 2a), indicating that stargazin enhances GluR1 function independent of trafficking. The TARP homologue γ -5 was inactive (Fig. 2a).

We next compared the effects of stargazin on responses to the partial agonist kainate. In oocytes, co-injecting stargazin with 0.1 ng GluR1 cRNA increased receptor surface expression by about 10-fold (Fig. 2b) but increased the kainate-evoked response by roughly 70-fold (Fig. 2b). In GluR1-injected oocytes, the absolute magnitude of kainate-evoked current was less than that of glutamate-evoked current (Fig. 2c). However, co-injection of stargazin with GluR1 greatly increased the efficacy of kainate (Fig. 2c, d). Stargazin potentiated kainate-evoked responses to both 'flip' and 'flop' alternatively spliced versions of GluR1 and GluR1–GluR2 heteromers (Fig. 2d). In outside-out patches from hippocampal neurons, AMPAR-mediated steady-state currents elicited by kainate were much larger than those elicited by glutamate (Fig. 2d), which resembles the results with stargazin in expression systems. In addition to stargazin, three related TARPs (γ -3, γ -4 and γ -8) all increased the relative efficacy of kainate (Fig. 2e), whereas γ -5, a homologue that does not traffic AMPARs, was inactive. The effects of stargazin and γ -3 were greater in magnitude than those of γ -4 and γ -8 (Fig. 2e). We constructed glutamate dose-response curves of peak current amplitude from HEK293 cells expressing GluR1 alone or with stargazin. Stargazin decreased the effective concentration for

¹Department of Physiology, and ²Department of Cellular and Molecular Pharmacology, University of California at San Francisco, San Francisco, California 94143, USA.

³Department of Pharmacology, Yale University School of Medicine, New Haven, Connecticut 06520, USA. ⁴Department of Degenerative Neurological Diseases, National Institute of Neuroscience, National Center of Neurology and Psychiatry, Kodaira, Tokyo 187-8502, Japan.

*These authors contributed equally to this work.

half-maximum response (EC_{50}) for glutamate from $57 \pm 9 \mu\text{M}$ ($n = 6$) to $24 \pm 6 \mu\text{M}$ ($n = 5$) (Fig. 2f). Finally, we found that stargazin does not change the current–voltage relationship for GluR1-flip or GluR1-flip–GluR2-flip AMPARs (Supplementary Fig. 1).

Distinct stargazin domains control AMPAR function

As our previous work showed that both extracellular and intracellular determinants in stargazin interact with AMPARs¹⁹, we asked whether these regions might differentially control AMPAR channel properties and trafficking. The ratio of kainate- and glutamate-evoked currents (relative kainate sensitivity) now provided a simple assay independent of absolute currents to evaluate the effect of stargazin on AMPAR channel properties. To exploit this, we constructed stargazin mutants in which specific domains were replaced by those from γ -5, which is structurally similar to stargazin¹⁵ but does not enhance glutamate-evoked currents (Fig. 3a). AMPAR preference for kainate was abolished by mutating the extracellular loop of stargazin (that is, replacing it with that from γ -5), even though this change did not interfere with receptor trafficking (Fig. 3a, Ex1). Conversely, replacing the cytoplasmic domain with that of γ -5 prevented stargazin's enhancement of receptor trafficking but preserved relative kainate selectivity (Fig. 3a, Cyto). To determine whether these domains are sufficient for stargazin's effects on AMPARs, we made the converse chimaeras by replacing domains of γ -5 with those from stargazin

(Fig. 3b). Restoring kainate sensitivity required the first extracellular and the second transmembrane domain of stargazin (Fig. 3b, 2Ex1TM2). Replacing just the cytoplasmic domain of γ -5 with that from stargazin selectively rescues receptor trafficking but not relative kainate sensitivity (Fig. 3b, 2Cyto). Finally, combining these two mutations in γ -5 reconstitutes both receptor trafficking and relative kainate sensitivity (Fig. 3b, 2Ex1TM2Cyto), demonstrating that these domains are necessary and sufficient for stargazin's actions. These data indicate that stargazin's effects on AMPAR agonist efficacy and trafficking are separable: the first extracellular loop of stargazin controls channel properties, whereas stargazin's cytoplasmic tail mediates receptor trafficking.

Stargazin ectodomain controls AMPAR gating

To explore the effects of stargazin on AMPAR desensitization, we used a fast glutamate perfusion system. Stargazin slowed GluR1 desensitization and deactivation of both GluR1 flip and flop forms in oocyte membrane patches (Fig. 4a–d), and these effects required the first extracellular loop of stargazin (see Ex1 in Fig. 4a–d). Stargazin also slowed desensitization of GluR4 in patches from transfected tsA201 cells (Fig. 5a, b). GluR4 produces larger steady-state currents than GluR1 (ref. 21), and this allowed us to quantify steady-state glutamate-evoked currents. Stargazin quadrupled the relative size of the steady-state current (GluR4 alone: $0.8 \pm 0.1\%$; GluR4 plus stargazin: $3.3 \pm 0.9\%$ of the peak current amplitude; $n = 6$ and 3, respectively; $P < 0.05$). Cyclothiazide (CTZ), which blocks AMPAR desensitization, similarly potentiated kainate- and glutamate-evoked peak currents in cells expressing GluR4 with stargazin (glutamate/kainate ratio without CTZ = 1.9 ± 0.4 , $n = 3$; with CTZ = 1.9 ± 0.1 ; $n = 3$).

To characterize further stargazin's effects, we recorded single-channel activity in patches from GluR4-transfected tsA201 cells. The patches gave peak currents of 20–40 pA (at -100 mV) and probably contained 10–20 channels. Single-channel openings could be distinguished late in the applications, and stargazin increased the frequency of these openings (Fig. 5c, d). Conductance levels and apparent open and shut times were analysed in records containing single-channel currents ($n = 6$ patches). Four discrete open levels were seen in all records. Without stargazin, most openings were to conductance levels of 9 and 20 pS (8.7 ± 0.7 pS, 19.5 ± 1.2 pS), with rare openings to larger levels (31 ± 1.2 pS and 44.8 ± 2.6 pS). Stargazin did not significantly alter the conductance levels but increased the frequency of openings to the largest conductance level by approximately sevenfold (Fig. 5e). The two smallest levels seen here are similar to the largest levels seen in some previous work on recombinant AMPARs^{22–24}, and levels close to our third level have been seen for GluR1 and GluR2 channels^{25,26}. Our largest open level has not been seen in previous studies of recombinant channels lacking stargazin; however, a similar large level was found in work on native receptors^{27,28}. Although stargazin did not alter open times or durations of brief shuttings significantly, it prolonged channel bursts (Table 1 and Fig. 5d, f; see Methods for the definition of bursts). The stargazin-mediated increases in large conductance openings (Fig. 5e) and burst lengths (Table 1) provide an explanation for the potentiation of steady-state currents (Figs 2 and 5a, b).

Several mechanisms might explain the effects of stargazin on channel kinetics. Delayed glutamate dissociation or altered rates of

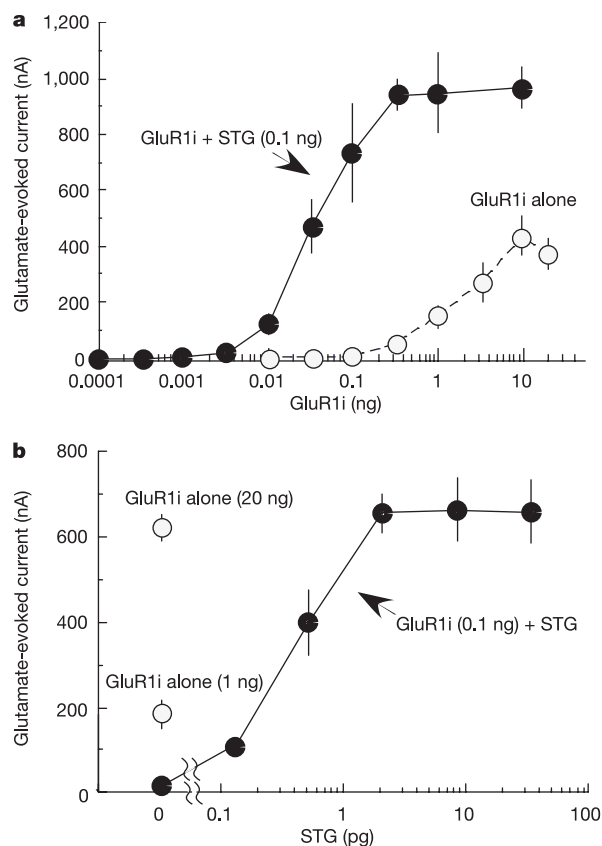


Figure 1 | Stargazin enhances glutamate-evoked currents in *Xenopus laevis* oocytes injected with GluR1. Oocytes were injected with GluR1-flip (GluR1i) and stargazin (STG) cRNAs, as indicated, and responses to $5 \mu\text{M}$ glutamate plus $50 \mu\text{M}$ cyclothiazide were recorded. **a**, Addition of stargazin cRNA enhances currents in oocytes co-injected with varying amounts of GluR1 cRNA. **b**, Injection of GluR1 cRNA in large amounts (1 or 20 ng) yields robust currents whereas injection of 0.1 ng GluR1 cRNA produces minimal glutamate-evoked currents. Addition of stargazin cRNA dose-dependently increases currents in oocytes co-injected with 0.1 ng GluR1 cRNA. Points show mean $\pm$ s.e.m. from six experiments.

Table 1 | Effect of stargazin on GluR4 burst length

Transfection	τ_{fast} (ms)	Area (%)	τ_{slow} (ms)	Area (%)
GluR4	1.1 ± 0.4	24 ± 10	3.8 ± 0.2	76 ± 10
GluR4 plus STG	3.6 ± 0.3	60 ± 4	12.3 ± 0.5	40 ± 4.0

The results are from three patches each with and without stargazin (STG) (τ_{fast} , $P < 0.05$; τ_{slow} , $P < 0.01$). Bursts were defined and distributions were fitted as described in the Methods.

channel gating might account for the slower deactivation. Delayed entry into the channel desensitized state would both slow current decay and increase the duration of bursts. Of these possible mechanisms, only an increased rate of channel opening would slow both deactivation and desensitization, increasing both burst duration and amplitude of steady-state currents (with no effect on apparent open times). Therefore, we conclude that stargazin alters AMPAR kinetics by increasing the rate constant for channel opening.

Stargazin ectodomain regulates synaptic transmission

To determine whether the stargazin-mediated effects on AMPAR biophysics have physiological relevance, we took advantage of the Ex1 stargazin- γ -5 chimaera that mediates receptor trafficking but does not slow receptor desensitization and deactivation (Fig. 4c, d). We reasoned that this chimaera, when expressed in neurons, would replace the endogenous TARP and disrupt its effects on channel properties. As a control, we found that overexpression of wild-type stargazin had no effect on the amplitude or decay kinetics of miniature excitatory postsynaptic currents (mEPSCs) (see Supplementary Fig. 2), which is consistent with previous findings^{29,30}. In contrast, overexpression of the Ex1 chimaera accelerated the decay of mEPSCs (Fig. 6a, b) and also decreased the peak amplitude of these currents (Fig. 6a, c). The reduction in mEPSC amplitude is an underestimate of the true effect because in Ex1-infected cells many more events fell below the detection threshold, as demonstrated by the decrease in apparent frequency (inter-event intervals for uninfected cells = 1.7 ± 0.3 s, $n = 17$; Ex1-infected cells = 4.6 ± 1.1 s, $n = 13$; $P < 0.05$).

To quantify the effect of Ex1 overexpression on AMPAR-mediated synaptic currents more accurately, we compared hippocampal evoked EPSCs in an infected cell and a neighbouring control cell. The chimaera decreased both the peak amplitude of the AMPAR EPSC (Fig. 6d, e) and the charge transfer mediated by AMPARs (uninfected = 2.2 ± 0.8 pC, $n = 8$; Ex1-infected = 0.7 ± 0.2 pC, $n = 8$; $P < 0.05$; data not shown). Although we detected no obvious difference in the decay of the evoked EPSCs, in both groups the kinetics of the evoked EPSC (uninfected: rise time = 7.8 ± 1.3 ms, decay = 11.5 ± 1.4 ms, $n = 9$; Ex1-infected: rise time = 8.4 ± 0.8 ms, decay = 11.5 ± 0.9 ms, $n = 9$) were considerably slower than those for the mEPSC (uninfected: rise time = 2.6 ± 0.2 ms, decay = 6.8 ± 0.7 ms, $n = 14$; Ex1-infected: rise time = 2.1 ± 0.2 ms, decay = 5.3 ± 0.3 ms, $n = 10$), presumably due to asynchronous release. As a control, the Ex1 chimaera had no effect on the NMDAR component of the EPSC in these experiments (Fig. 6d, f). In a final series of experiments we performed infections to 'rescue' AMPARs in cultured cerebellar granule cells from stargazer mice. As compared to stargazin transfections, neurons transfected with the Ex1 mutant had EPSCs with faster kinetics and smaller amplitudes (Supplementary Fig. 3). These results indicate that the ectodomain of stargazin controls the strength and kinetics of excitatory synaptic transmission in the central nervous system.

Discussion

This study demonstrates that stargazin not only mediates AMPAR trafficking, but also controls the channel's biophysical properties. Analysis of mutants shows that the stargazin carboxy-terminal

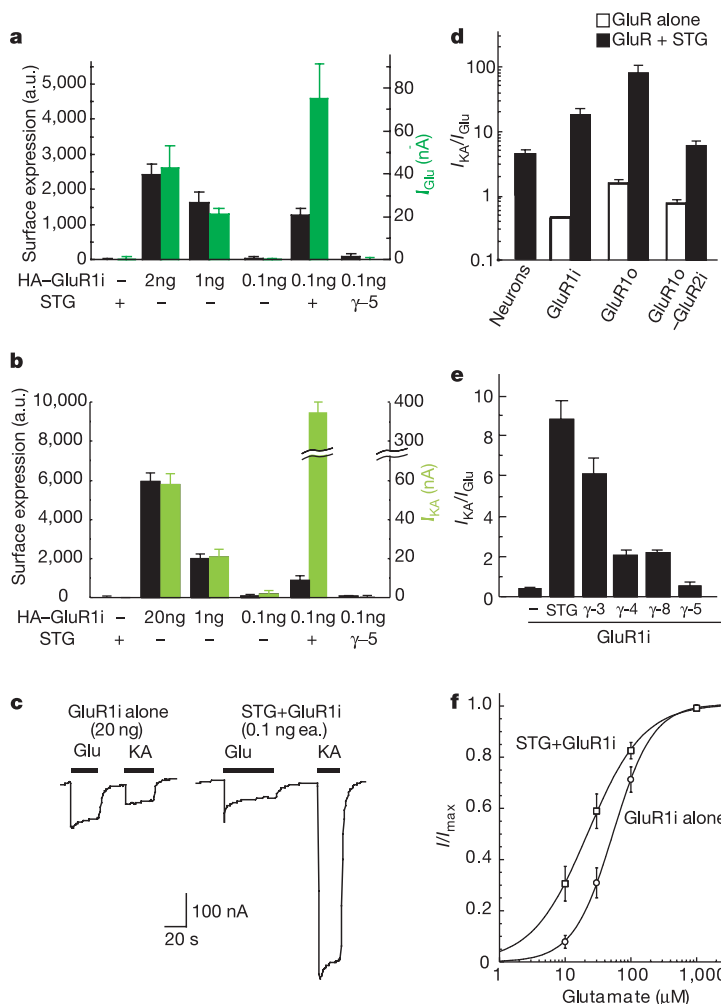


Figure 2 | Stargazin mediates AMPAR trafficking and also modulates AMPAR agonist efficacy. **a, b**, Oocytes were injected with HA-tagged GluR1-flip (HA-GluR1i) and stargazin (STG) or γ -5 cRNA as indicated. Receptor surface expression and currents evoked by glutamate (10 μ M; I_{Glu}) or kainate (40 μ M; I_{KA}) were quantified in parallel by chemiluminescence ($n = 9$) and TEVC recordings ($n = 5$), respectively. a.u., arbitrary unit. Stargazin enhances surface expression of limiting amounts (0.1 ng) of HA-GluR1 but has a much greater effect on glutamate- (**a**) and kainate-evoked (**b**) currents. **c, d**, Oocytes injected with GluR1-flip alone (GluR1i) are more sensitive to glutamate (500 μ M) than to kainate (500 μ M) ($n = 4$). Co-injection of stargazin (STG) with GluR1i enhances kainate sensitivity in oocytes ($n = 4$). **d**, Stargazin also enhances kainate sensitivity for homomeric GluR1 flop (GluR1o) and heteromeric (GluR1o-GluR2i) AMPARs. In outside-out patches from hippocampal neurons, kainate produces larger AMPAR-mediated (GYKI-53655 sensitive) currents than does glutamate ($n = 6$). **e**, Stargazin and the other TARPs (γ -3, γ -4 and γ -8) all increase kainate efficacy at GluR1 whereas γ -5 does not. **f**, Dose-response curves of peak current amplitude show that stargazin increases the affinity of GluR1 to glutamate in HEK293 cells. Data in **a, b, d-f** show mean $\pm$ s.e.m. from the indicated number of experiments.

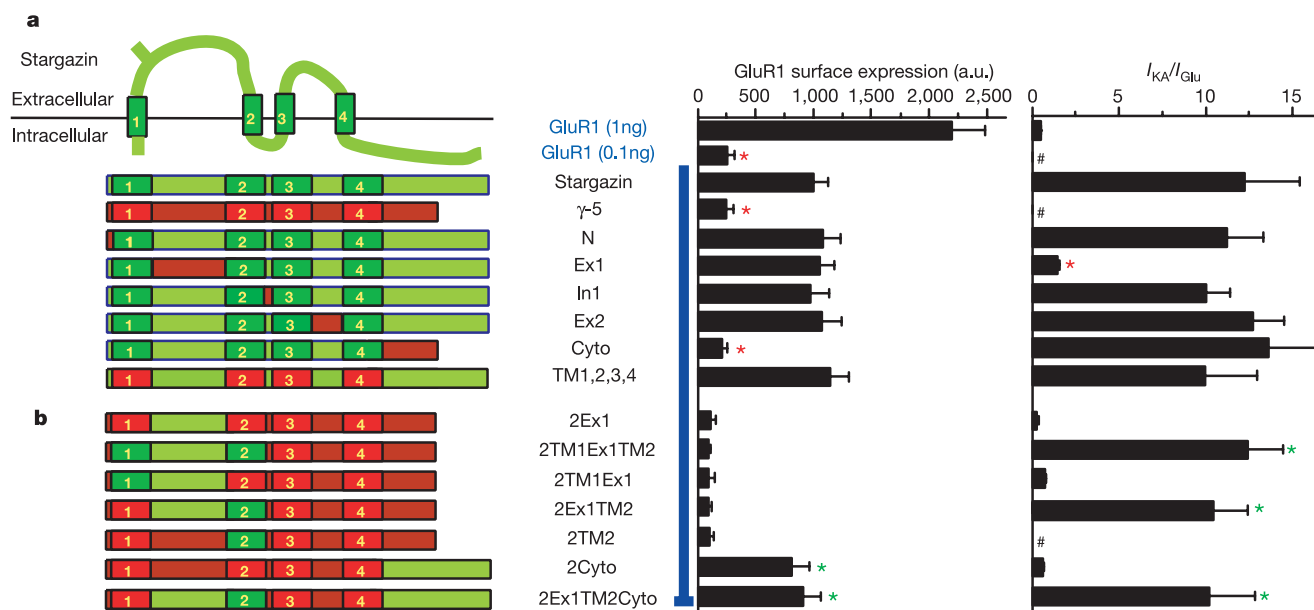


Figure 3 | Stargazin regulates AMPAR trafficking and agonist efficacy by distinct mechanisms. Oocytes were injected with 1 ng or 0.1 ng GluR1 cRNA and stargazin- γ -5 chimaeric constructs as indicated. Receptor surface expression and kainate/glutamate-evoked currents (I_{KA}/I_{Glu}) were quantified in parallel by chemiluminescence and TEVC recordings, respectively (shown as mean $\pm$ s.e.m.). **a**, Stargazin enhances receptor trafficking and produces kainate-preferring channels, whereas γ -5 is inactive. Replacing the stargazin first extracellular domain with that of γ -5 (Ex1) does not disrupt surface expression but selectively diminishes kainate responses. On the other hand, replacing the cytoplasmic domain of stargazin (Cyto) maintains kainate selectivity but abolishes enhanced receptor

trafficking. Red asterisks indicate a significant difference compared with GluR1 plus STG; $P < 0.05$. **b**, Swapping domains of γ -5 with those of stargazin shows that the stargazin first extracellular and second transmembrane domains (2Ex1TM2) are sufficient to induce kainate-selective responses. The stargazin cytoplasmic tail (2Cyto) is sufficient to induce receptor trafficking. Combining these regions (2Ex1TM2Cyto) is sufficient to restore both trafficking and pharmacological influences of stargazin. The hash symbol indicates that kainate-induced currents were undetectable. Green asterisks indicate a significant difference compared with GluR1 alone; $P < 0.05$.

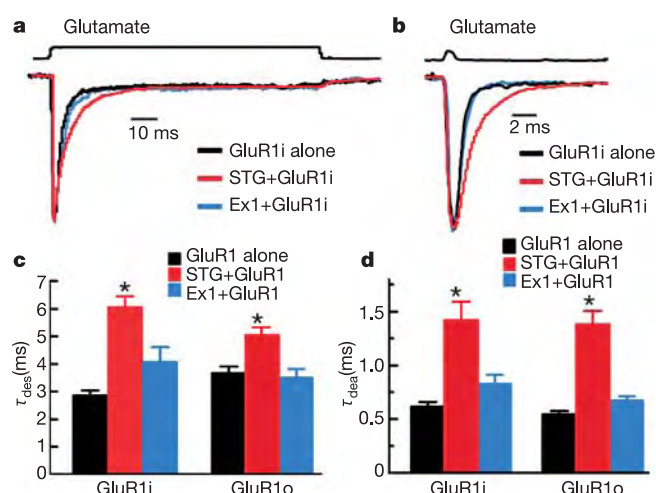


Figure 4 | Stargazin slows AMPAR desensitization and deactivation. **a, b**, Superimposed responses of outside-out oocyte patches to 100-ms (**a**) or 1-ms (**b**) applications of glutamate (10 mM) for GluR1i alone (GluR1i; black), GluR1i with stargazin (STG+GluR1i; red) or GluR1i with Ex1, the γ -5 first extracellular domain in stargazin (Ex1+GluR1i; blue). Open-tip responses recorded at the end of experiments are shown above. Responses are normalized to their peak amplitude to compare their time course. **c, d**, Stargazin slowed desensitization (**c**; τ_{des} ; $n = 6$) and deactivation (**d**; τ_{des} ; $n = 5$) for GluR1i (GluR1i) and flop (GluR1o), whereas Ex1 did not. The asterisk indicates a significant difference compared with GluR1 alone; $P < 0.01$. All analysed traces show peak amplitudes between 20 and 400 pA. Data in **c, d** show mean $\pm$ s.e.m. for the indicated number of experiments.

tail dictates AMPAR trafficking, whereas the ectodomain of stargazin influences channel gating. Disrupting stargazin ectodomain interactions alters the size and shape of hippocampal EPSCs, indicating that stargazin serves as a functional AMPAR auxiliary subunit at the synapse. The discovery of an auxiliary subunit that controls the gating of a mammalian ionotropic receptor is unprecedented; however, the transmembrane protein SOL-1 may similarly function as a necessary subunit for the glutamate-gated channel GLR-1 in *Caenorhabditis elegans*³¹.

Mechanisms for stargazin control of AMPARs

The role of the stargazin C-terminal tail in trafficking suggests that interactions of stargazin and AMPAR subunits with cytoplasmic proteins regulate receptor surface expression. Numerous cytosolic proteins have been shown to bind to AMPARs (for a review, see refs 6–10), but these cytoplasmic partners all show specificity for interaction with specific AMPAR subunits. By contrast, stargazin traffics all AMPAR subunits¹⁹, so the actions of stargazin are probably independent of these previously reported proteins.

Structural studies of the isolated GluR ligand-binding core indicate that agonists induce closure of the clamshell-shaped binding site and that this movement gates the channel. Whereas glutamate induces a large conformational change and serves as a full agonist, kainate causes much less movement and acts as a partial agonist³². Analysis of a series of partial agonists showed that the GluR2 ligand-binding core takes on a range of conformational states that control the probability of opening to discrete sub-conductance states²⁴. That stargazin makes kainate a more efficacious agonist suggests that stargazin's interactions with the ligand-binding core enhance kainate-induced domain closure. This proposal may explain why kainate evokes only small-conductance openings of heterologously expressed GluR4 (ref. 22), whereas it produces much larger

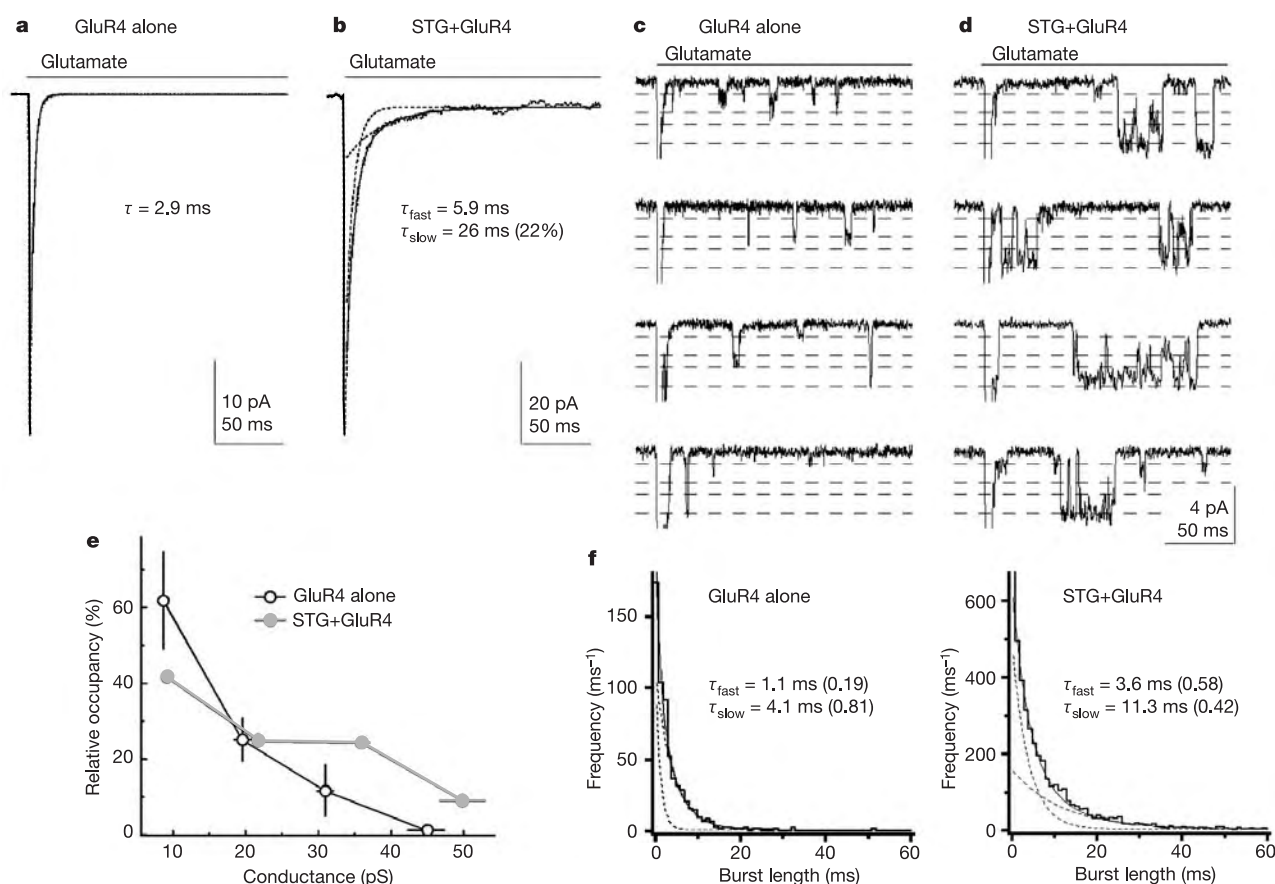


Figure 5 | Stargazin increases AMPAR open channel probability.

a, b, Inward currents were averaged from 50 applications of glutamate (10 mM) to patches from transfected cells ($n = 6$ for GluR4 alone; $n = 3$ for STG plus GluR4). The current decay of receptors with stargazin contains two exponential components (dotted lines) with time constants that are slower than the single time constant seen with GluR4 alone. **c, d**, Examples of unitary inward currents seen during 200-ms applications of glutamate (10 mM) (**c**, 500 applications; **d**, 1,000 applications). Channel activations are longer in receptors with stargazin. The dotted lines indicate the four open levels detected in each patch. Peak currents were 30–40 pA and are off-scale. **e**, Plot of the mean percentage of time spent at each of the four open levels

for GluR4 alone (open circles; black) and GluR4 with stargazin (filled circles; grey). Error bars indicate s.e.m. for the relative occupancies and conductance levels ($n = 3$). **f**, Histograms of burst-length distributions from a patch containing GluR4 (left, 610 bursts) and a patch containing GluR4 and stargazin (right, 4,464 bursts). The distributions were fitted (solid line) with two exponential components (dotted lines) (τ_{fast} ; $P < 0.05$, τ_{slow} ; $P < 0.01$). Mean burst length (s.e.m. in parentheses) is indicated for the fast and slow distributions. Note that long bursts are more common with stargazin. The first bin in the stargazin histogram contained 1,012 bursts and is off-scale (see Methods). One burst from the GluR4 patch and 34 bursts from the stargazin patch had durations above 60 ms (off-scale, right).

conductance openings of GluR4 in cerebellar granule cells, which express stargazin²⁸. Structural studies of stargazin–AMPA complexes will be necessary to address this issue directly.

Our finding that stargazin increases glutamate-induced large-conductance openings and burst length suggests that stargazin further increases the efficacy for glutamate. These single-channel results, together with the effects of stargazin on desensitization and deactivation, indicate that stargazin's augmentation of burst length reflects an increase in the rate constant for channel opening. As predicted by the slowed deactivation, stargazin increased receptor affinity, but this modest change in affinity cannot explain the massive influences of stargazin on AMPAR function. The molecular basis for these effects will also require future structural studies of the stargazin–AMPA complex.

Functional significance of stargazin–AMPA interactions

Action potential firing in neurons of the central nervous system is determined by the magnitude and coincidence of synaptic inputs, and depends in large part on synaptic AMPAR number and properties. Previously, we found that stargazin's interaction with PSD-95 controls synaptic AMPAR number³⁰. Here, we demonstrate that stargazin also modulates the biophysical properties of AMPAR

channels, increasing the magnitude and slowing the decay of AMPAR-mediated synaptic currents, which we estimate results in a threefold increase in the net charge transfer for a given quantum of synaptically released glutamate. These effects of stargazin, as well as the established effects of distinct AMPAR isoforms and splice variants^{33,34}, influence neuronal firing in response to synaptic input. The differential expression of TARP isoforms in discrete neuronal populations¹⁵ may contribute to synapse-specific timing, as different isoforms differ in their ability to modify AMPAR function (see Fig. 2e). Future studies of recombinant GluR subunits will require inclusion of appropriate TARPs to model the kinetics of AMPARs *in vivo*.

This work also has relevance for the therapeutic pharmacology of AMPARs. Over-activation of AMPARs may contribute to neurodegenerative diseases³⁵. Modulation of the TARP–AMPA interaction offers a novel target for downregulating the receptors. On the other hand, reagents that limit AMPAR desensitization—AMPA potentiators—show promise as cognitive enhancers^{36,37}, and may also be effective in treating schizophrenia and other mental disorders^{38,39}. Regulation of the ectodomain interaction of TARPs with AMPARs should provide an approach for these neurological disorders.

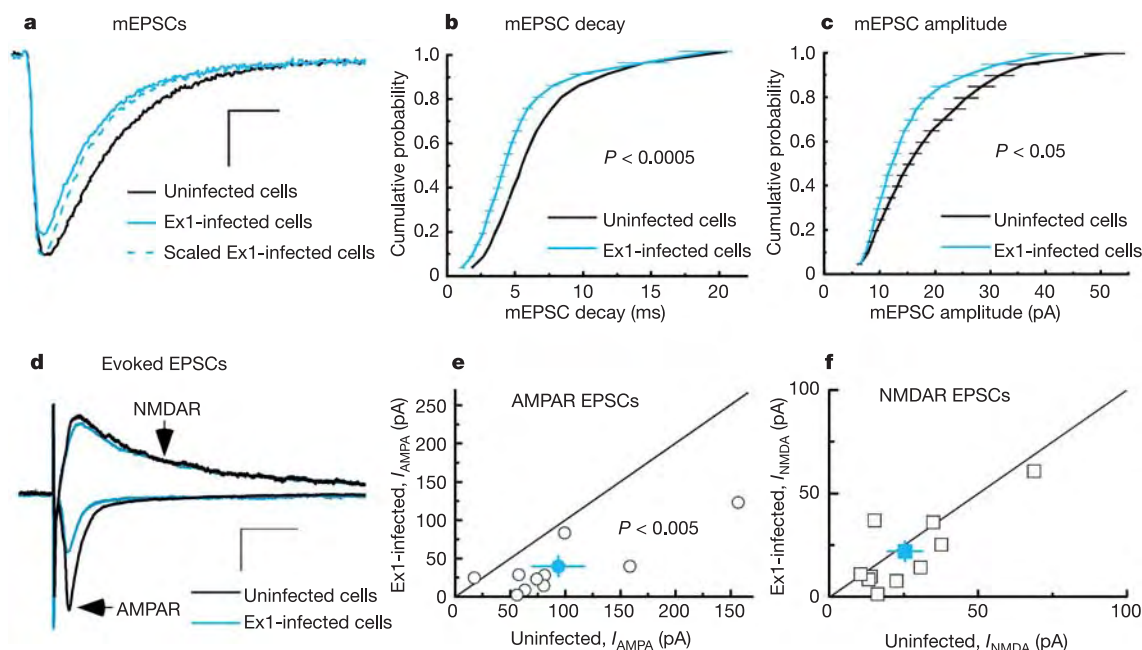


Figure 6 | The ectodomain of stargazin influences the amplitude and decay of EPSCs in hippocampal neurons. **a**, Miniature EPSCs were recorded (in the presence of TTX) from hippocampal CA3 pyramidal cells in slice culture from uninfected (black) or infected (blue) cells expressing the Ex1 chimera (γ -5 first extracellular domain in stargazin). The traces are representative averages of more than 400 individual events in five control and five Ex1-infected cells. Scale bar: 5 pA, 5 ms. **b**, Synaptic events were fitted to exponential mEPSC decay times. Cumulative frequency distribution shows that the decay in infected neurons is significantly faster than that in control uninfected cells ($n = 10$ infected and $n = 14$ for control; $P < 0.0005$). **c**, Cumulative frequency distribution shows that the amplitude in Ex1-infected neurons is smaller than that in control uninfected neurons

($P < 0.05$). **d**, Evoked EPSCs were recorded in hippocampal CA3 pyramidal cells from uninfected (black) and infected (blue) neurons expressing the Ex1 chimera. Scale bar: 100 pA, 40 ms. **e**, The size of the AMPAR EPSC (-60 mV) in uninfected cells was compared with the size of the AMPAR EPSC in neighbouring Ex1-infected cells (ten pairs). Note that all but one of the points falls to the right of the line of unity, indicating a substantial decrease in the AMPAR EPSC in Ex1-infected cells. **f**, The NMDAR receptor component was measured at $+40$ mV and at 100-ms latency, by which time the AMPAR component has decayed to zero. Expression of Ex1 has no effect on the NMDAR EPSC. Points in **e**, **f** show mean $\pm$ s.e.m. from ten pairs of experiments.

Note added in proof: While this paper was under review, a paper by A. Priel *et al.* appeared that confirms some of our findings⁴⁰.

METHODS

Electrophysiology using *Xenopus laevis* oocytes. Two-electrode voltage clamp (TEVC) recordings were performed as described¹⁹. Briefly, GluR1, stargazin and γ -5 constructs were subcloned into pGEM-HE vector and cRNAs were transcribed *in vitro* using T7 mMessage mMachine (Ambion). TEVC analysis was performed 2 days after injection at room temperature in recording solution containing (in mM): 90 NaCl, 1.0 KCl, 1.5 CaCl₂ and 10 HEPES (pH 7.4). The membrane potential was held at -70 mV.

Surface labelling of oocytes. Surface labelling was performed as described²⁰. Briefly, oocytes 3 days after injection were incubated for 1 h with $0.25 \mu\text{g ml}^{-1}$ rat anti-HA antibody (3F10, Roche) followed by 30 min with horseradish-peroxidase-conjugated anti-rat immunoglobulin. Individual oocytes were then placed into $50 \mu\text{l}$ of SuperSignal ELISA Femto maximum sensitivity substrate (Pierce) and quantified by chemiluminescence using a TD20/20 Luminometer (Turner Designs).

Outside-out patch recordings. Outside-out patches from injected oocytes were obtained as described previously⁴¹. Outside-out patch recording was carried out using an EPC-8 amplifier (HEKA) under continuous perfusion with frog Ringer's solution⁴¹. The patch pipette was prepared from borosilicate glass capillaries (WPI) and had $4\text{--}7\text{ M}\Omega$ input resistance⁴¹. Responses were filtered at 10 kHz and digitized at $26 \mu\text{s}$ per point. The holding potential was maintained at -60 mV. Fast application of glutamate was performed using methods described previously⁴¹. Briefly, glutamate (10 mM) was applied by perfusion of the patch membrane with theta tubes driven by a piezo manipulator (Burleigh PZ-150M). After recording, the patch membrane was blown off with positive pressure and the junction current between the control solution and 10% frog Ringer's solution was measured to monitor solution exchange without moving the patch pipette and the theta tube. Responses to glutamate having a 20–80% rise time less than $400 \mu\text{s}$ were used for analysis. The decay phase of the response was fitted to single exponential functions using Igor Pro (WaveMetrics). The

decay time constant was calculated by fitting the decay of the response with a single exponential function.

Patch clamp recording and virus infection. Hippocampal slice cultures were prepared from 6–11-day-old rats, as described⁴². Stargazin-Ex1 or wild-type stargazin were co-expressed with GFP 4–6 days later with Semliki forest virus by using an internal ribosome entry site (IRES) construct. Recordings were made from infected cells 1–2 days after infection, using $2\text{--}3\text{ M}\Omega$ glass electrodes. The internal and external solutions were as described previously⁴². mEPSCs, collected in the presence of tetrodotoxin (TTX), were automatically detected with custom software, and each event was fitted with a mono-exponential decay function. Cumulative frequency distributions were analysed by Kolmogorov–Smirnov tests. Evoked EPSCs were recorded and analysed as described previously⁴², except that recordings from control and nearby infected cells were done sequentially. Paired recordings were analysed using paired Student's *t*-test. For all other statistical analysis, unpaired Student's *t*-test was used.

Single-channel recording and data analysis. Patch-clamp recordings were performed 12–24 h after transfection at room temperature with an EPC 9 amplifier, as previously described⁴³. All recordings were from excised outside-out patches held at -100 mV. Glutamate (10 mM) was added to the external solution and was applied with theta pipettes mounted on a piezoelectric bimorph, as described previously²¹. The rate of solution exchange estimated from open-tip potentials was $100\text{--}200 \mu\text{s}$.

Glutamate-evoked currents were analogue low-pass filtered at 10 kHz, sampled at 40 kHz and written directly to the hard drive of two computers. One computer stored the records in PULSE format (the software used to run the EPC 9) and the other stored the data in QuB format (the software used for single-channel analysis). For analysis of macroscopic current decays, the PULSE records were exported to IGOR, averaged and fitted with single- or bi-exponential functions, as described⁴³. For analysis of unitary currents, the data were digitally low-pass filtered at 2 kHz. The data were edited manually in QuB to isolate unitary events that occurred late in the glutamate applications, and the resulting record was idealized with the SKM algorithm in QuB (using hidden Markov models) to identify single-channel transitions and estimate conductance levels.

Given the multiple open levels present, there was no single time resolution that applied to all possible types of transitions. The resolution was set to 100 μ s to avoid missing brief, large-amplitude events. After inspecting the idealized record and removing dubious events, mean open and shut times were obtained with log-binned fitting of the dwell-time distributions for each conductance level (using maximum interval likelihood subroutines in QuB). The shut time distributions with and without stargazin were similar, and bursts of openings were defined as a series of openings separated by shuttings shorter than a critical time (2.5–3 ms for the six patches). Distributions of the durations of bursts were fitted in QuB with two exponential components. In each stargazin record, there was an excess of brief bursts (presumably single openings) that represented less than 10% of the total number of bursts. Because of the small number and brevity of these bursts, the time constant and relative area of this additional component were poorly defined, and they were not included in the results reported here.

Received 8 November 2004; accepted 12 April 2005.

Published online 27 April 2005.

- Nakanishi, S. Molecular diversity of glutamate receptors and implications for brain function. *Science* **258**, 597–603 (1992).
- Wisden, W. & Seeburg, P. H. Mammalian ionotropic glutamate receptors. *Curr. Opin. Neurobiol.* **3**, 291–298 (1993).
- Hollmann, M. & Heinemann, S. Cloned glutamate receptors. *Annu. Rev. Neurosci.* **17**, 31–108 (1994).
- Mayer, M. L. & Armstrong, N. Structure and function of glutamate receptor ion channels. *Annu. Rev. Physiol.* **66**, 161–181 (2004).
- Dingledine, R., Borges, K., Bowie, D. & Traynelis, S. F. The glutamate receptor ion channels. *Pharmacol. Rev.* **51**, 7–61 (1999).
- Sheng, M. & Kim, M. J. Postsynaptic signaling and plasticity mechanisms. *Science* **298**, 776–780 (2002).
- Malinow, R. & Malenka, R. C. AMPA receptor trafficking and synaptic plasticity. *Annu. Rev. Neurosci.* **25**, 103–126 (2002).
- Bredt, D. S. & Nicoll, R. A. AMPA receptor trafficking at excitatory synapses. *Neuron* **40**, 361–379 (2003).
- Song, I. & Haganir, R. L. Regulation of AMPA receptors during synaptic plasticity. *Trends Neurosci.* **25**, 578–588 (2002).
- Barry, M. F. & Ziff, E. B. Receptor trafficking and the plasticity of excitatory synapses. *Curr. Opin. Neurobiol.* **12**, 279–286 (2002).
- Chen, L. *et al.* Stargazin regulates synaptic targeting of AMPA receptors by two distinct mechanisms. *Nature* **408**, 936–943 (2000).
- Hashimoto, K. *et al.* Impairment of AMPA receptor function in cerebellar granule cells of ataxic mutant mouse stargazer. *J. Neurosci.* **19**, 6027–6036 (1999).
- Letts, V. A. *et al.* The mouse stargazer gene encodes a neuronal Ca^{2+} -channel γ subunit. *Nature Genet.* **19**, 340–347 (1998).
- Chen, L., Bao, S., Qiao, X. & Thompson, R. F. Impaired cerebellar synapse maturation in waggler, a mutant mouse with a disrupted neuronal calcium channel γ subunit. *Proc. Natl Acad. Sci. USA* **96**, 12132–12137 (1999).
- Tomita, S. *et al.* Functional studies and distribution define a family of transmembrane AMPA receptor regulatory proteins. *J. Cell Biol.* **161**, 805–816 (2003).
- Catterall, W. A. Structure and regulation of voltage-gated Ca^{2+} channels. *Annu. Rev. Cell Dev. Biol.* **16**, 521–555 (2000).
- Arikath, J. & Campbell, K. P. Auxiliary subunits: essential components of the voltage-gated calcium channel complex. *Curr. Opin. Neurobiol.* **13**, 298–307 (2003).
- Chen, L., El-Husseini, A., Tomita, S., Bredt, D. S. & Nicoll, R. A. Stargazin differentially controls the trafficking of α -amino-3-hydroxyl-5-methyl-4-isoxazolepropionate and kainate receptors. *Mol. Pharmacol.* **64**, 703–706 (2003).
- Tomita, S., Fukata, M., Nicoll, R. A. & Bredt, D. S. Dynamic interaction of stargazin-like TARPs with cycling AMPA receptors at synapses. *Science* **303**, 1508–1511 (2004).
- Zerangue, N., Schwappach, B., Jan, Y. N. & Jan, L. Y. A new ER trafficking signal regulates the subunit stoichiometry of plasma membrane K(ATP) channels. *Neuron* **22**, 537–548 (1999).
- Robert, A. & Howe, J. R. How AMPA receptor desensitization depends on receptor occupancy. *J. Neurosci.* **23**, 847–858 (2003).
- Swanson, G. T., Kamboj, S. K. & Cull-Candy, S. G. Single-channel properties of recombinant AMPA receptors depend on RNA editing, splice variation, and subunit composition. *J. Neurosci.* **17**, 58–69 (1997).
- Banke, T. G. *et al.* Control of GluR1 AMPA receptor function by cAMP-dependent protein kinase. *J. Neurosci.* **20**, 89–102 (2000).
- Jin, R., Banke, T. G., Mayer, M. L., Traynelis, S. F. & Gouaux, E. Structural basis for partial agonist action at ionotropic glutamate receptors. *Nature Neurosci.* **6**, 803–810 (2003).
- Derkach, V., Barria, A. & Soderling, T. R. Ca^{2+} /calmodulin-kinase II enhances channel conductance of α -amino-3-hydroxy-5-methyl-4-isoxazolepropionate type glutamate receptors. *Proc. Natl Acad. Sci. USA* **96**, 3269–3274 (1999).
- Mansour, M., Nagarajan, N., Nehring, R. B., Clements, J. D. & Rosenmund, C. Heteromeric AMPA receptors assemble with a preferred subunit stoichiometry and spatial arrangement. *Neuron* **32**, 841–853 (2001).
- Smith, T. C., Wang, L. Y. & Howe, J. R. Heterogeneous conductance levels of native AMPA receptors. *J. Neurosci.* **20**, 2073–2085 (2000).
- Wyllie, D. J., Traynelis, S. F. & Cull-Candy, S. G. Evidence for more than one type of non-NMDA receptor in outside-out patches from cerebellar granule cells of the rat. *J. Physiol. (Lond.)* **463**, 193–226 (1993).
- Tomita, S., Stein, V., Stocker, T. J., Nicoll, R. A. & Bredt, D. S. Bidirectional synaptic plasticity regulated by phosphorylation of stargazin-like TARPs. *Neuron* **45**, 269–277 (2005).
- Schnell, E. *et al.* Direct interactions between PSD-95 and stargazin control synaptic AMPA receptor number. *Proc. Natl Acad. Sci. USA* **99**, 13902–13907 (2002).
- Zheng, Y., Mellem, J. E., Brockie, P. J., Madsen, D. M. & Maricq, A. V. SOL-1 is a CUB-domain protein required for GLR-1 glutamate receptor function in *C. elegans*. *Nature* **427**, 451–457 (2004).
- Armstrong, N. & Gouaux, E. Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: crystal structures of the GluR2 ligand binding core. *Neuron* **28**, 165–181 (2000).
- Geiger, J. R. *et al.* Relative abundance of subunit mRNAs determines gating and Ca^{2+} permeability of AMPA receptors in principal neurons and interneurons in rat CNS. *Neuron* **15**, 193–204 (1995).
- Jonas, P. The time course of signaling at central glutamatergic synapses. *News Physiol. Sci.* **15**, 83–89 (2000).
- Koh, J. Y., Goldberg, M. P., Hartley, D. M. & Choi, D. W. Non-NMDA receptor-mediated neurotoxicity in cortical culture. *J. Neurosci.* **10**, 693–705 (1990).
- Staubli, U., Rogers, G. & Lynch, G. Facilitation of glutamate receptors enhances memory. *Proc. Natl Acad. Sci. USA* **91**, 777–781 (1994).
- Lynch, G. *et al.* Evidence that a positive modulator of AMPA-type glutamate receptors improves delayed recall in aged humans. *Exp. Neurol.* **145**, 89–92 (1997).
- Goff, D. C. *et al.* A placebo-controlled pilot study of the amphetamine CX516 added to clozapine in schizophrenia. *J. Clin. Psychopharmacol.* **21**, 484–487 (2001).
- O'Neill, M. J. *et al.* Neurotrophic actions of the novel AMPA receptor potentiator, LY404187, in rodent models of Parkinson's disease. *Eur. J. Pharmacol.* **486**, 163–174 (2004).
- Priel, A. *et al.* Stargazin reduces desensitization and slows deactivation of the AMPA-type glutamate receptors. *J. Neurosci.* **25**, 2682–2686 (2005).
- Sekiguchi, M., Nishikawa, K., Aoki, S. & Wada, K. A desensitization-selective potentiator of AMPA-type glutamate receptors. *Br. J. Pharmacol.* **136**, 1033–1041 (2002).
- Stein, V., House, D. R., Bredt, D. S. & Nicoll, R. A. Postsynaptic density-95 mimics and occludes hippocampal long-term potentiation and enhances long-term depression. *J. Neurosci.* **23**, 5503–5506 (2003).
- Robert, A., Irizarry, S. N., Hughes, T. E. & Howe, J. R. Subunit interactions and AMPA receptor desensitization. *J. Neurosci.* **21**, 5574–5586 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank K. Kam for help with analysing mEPSCs; A. Tzingounis for discussions; A. Sui for technical assistance; and R. Moberg for help with preparation of the manuscript. D.S.B. is supported by grants from the National Institutes of Health, the Christopher Reeve Paralysis Foundation and the Human Frontier Science Program. R.A.N. is supported by grants from NIH. J.R.H. is supported by grants from NIH. K.W. is supported by grants from Grants-in-Aid for Scientific Research from the Ministry of Health, Labour and Welfare of Japan, and Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan. D.S.B. is an established investigator of the American Heart Association. R.A.N. is a member of the Keck Center for Integrative Neuroscience and the Silvio Conte Center for Neuroscience Research. S.T. is supported by a grant from NIH. H.A. is a Howard Hughes Medical Institute predoctoral fellow.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.S.B. (bredt@itsa.ucsf.edu) or R.A.N. (nicoll@cnp.ucsf.edu).

A structural basis for allosteric control of DNA recombination by λ integrase

Tapan Biswas^{1*}, Hideki Aihara^{1*}, Marta Radman-Livaja², David Filman¹, Arthur Landy² & Tom Ellenberger¹

Site-specific DNA recombination is important for basic cellular functions including viral integration, control of gene expression, production of genetic diversity and segregation of newly replicated chromosomes, and is used by bacteriophage λ to integrate or excise its genome into and out of the host chromosome. λ recombination is carried out by the bacteriophage-encoded integrase protein (λ -int) together with accessory DNA sites and associated bending proteins that allow regulation in response to cell physiology. Here we report the crystal structures of λ -int in higher-order complexes with substrates and regulatory DNAs representing different intermediates along the reaction pathway. The structures show how the simultaneous binding of two separate domains of λ -int to DNA facilitates synapsis and can specify the order of DNA strand cleavage and exchange. An intertwined layer of amino-terminal domains bound to accessory (arm) DNAs shapes the recombination complex in a way that suggests how arm binding shifts the reaction equilibrium in favour of recombinant products.

λ -int catalyses an ordered, pair-wise exchange of four DNA strands between two different pairs of recombination substrates^{1,2}. During integration, λ -int aligns the bacteriophage attachment site *attP* with the bacterial attachment site *attB* and recombines these sequences to generate the recombination joints *attL* and *attR* flanking the integrated prophage (Fig. 1a, b). During the transition to lytic growth, the bacteriophage DNA is excised to regenerate *attP* and *attB*. In both reactions, the analogous pair of DNA strands ('top' strands) is exchanged first^{3,4} to form a branched, four-way DNA intermediate known as a Holliday junction. Subsequent exchange of 'bottom' strands resolves the Holliday junction into linear recombinant products². Although integration and excision might appear to be reciprocal reactions (Fig. 1), they involve different substrates and are effectively irreversible⁵. The recombination machinery is configured differently during integration or excision by two different overlapping subsets of accessory factors and binding sites that bend the DNA arms flanking the core sites of strand exchange^{2,6}. DNA bending is a prerequisite for the simultaneous interactions with arm and core sites^{6–9} that deliver λ -int to lower-affinity core sites¹⁰. Arm-binding interactions allosterically enhance the fidelity of DNA strand exchange¹¹ and bias the outcome of Holliday junction resolution in favour of the recombined products¹².

Two related tyrosine recombinases—the bacteriophage P1 Cre and yeast FLP proteins—lack the regulatory arm-binding functions of λ -int and catalyse a reversible exchange of DNA strands between identical sites^{13,14}. Crystal structures of Cre^{15,16} and FLP¹⁷ show that a protein tetramer mediates the synapse of two strongly bent antiparallel DNAs closely resembling a Holliday junction. These structures suggest that only a small shift in subunit packing is sufficient to redirect DNA cleavage and exchange activities from one pair of strands to the other, in order to resolve the Holliday junction into products^{13,14}. However, the molecular mechanism of isomerization is poorly understood.

In contrast to Cre and FLP, recombination by λ -int depends

absolutely on secondary interactions with the DNA arms flanking the site of strand exchange. Arm binding by λ -int provides a control point for site-specific recombination and facilitates crystallization of trapped intermediates from the multi-step reaction. We have determined crystal structures of λ -int tetramers complexed to arm and core DNAs in different conformations that provide a series of views along the recombination pathway (Fig. 1c–e). The two distinct DNA-binding activities of λ -int dictate a specific arrangement of core and arm sites that facilitates synapsis and explains the order of strand cleavage and exchange. Arm binding constrains the topology of the bound DNA and promotes isomerization towards a conformation that is proficient for the second exchange of DNA strands to complete recombination. This allosteric control makes λ recombination a conditional, effectively irreversible molecular switch during the bacteriophage lifecycle.

Assembly of a recombinogenic complex

λ -int was crystallized in three different complexes with either a Holliday junction¹⁸ or a COC' DNA substrate¹⁹ in the presence and absence of oligonucleotides containing tandem P'1–P'2 binding sites derived from *attP*¹¹ (Fig. 1). The crystal structures of these recombination complexes were determined by isomorphous replacement methods (see Methods and Supplementary Table S1), and they capture different steps of recombination. A synaptic complex consisting of an N-terminally truncated λ -int(75–356) tetramer bound to two COC' core sites represents an early step after the initial DNA cleavage but before strand exchange (Fig. 1c). In the complex with full-length λ -int and arm DNAs, the cleaved DNA strands have exchanged but ligation is prevented by a modification of the substrate DNA¹⁹ (Fig. 1d). The synaptic and post-exchange complexes crystallized in different space groups with a crystallographic two-fold axis relating two λ -int dimers and generating a two-fold symmetric tetramer. A Holliday junction intermediate (Fig. 1e) together with arm DNAs was crystallized in a complex with the λ -int(Tyr342Phe) mutant that is unable to cleave DNA and resolve the Holliday junction into

¹Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. ²Division of Biology and Medicine, Brown University, Providence, Rhode Island 02912, USA.

*These authors contributed equally to this work.

products. This complex has a roughly two-fold symmetry and is configured for the second strand exchange reaction.

The Holliday junction complex exemplifies the overall organization of λ recombination complexes. Upon DNA binding, the three functional domains of λ -int^{20,21} assemble into the three distinct layers of the tetramer. The N-terminal domain (N domain; residues 1–63) that binds to arm sites is connected to the core-binding domain (CB domain; residues 75–175) by a short, α -helical coupler (residues 64–74). The CB domain is, in turn, joined to the C-terminal catalytic domain (residues 176–356) through another extended segment

(residues 160–176; Fig. 2)¹⁹. All three domains of λ -int collaborate to engage the core and arm DNAs, establishing a tightly knit but flexible recombinogenic complex.

The λ -int tetramer has a cyclically permuted topology in which each N domain packs on top of the neighbouring CB domain (Fig. 2a). Fluorescence resonance energy transfer studies of similar complexes also show this permuted arrangement²². The N domains of two adjacent subunits adopt different conformations that orient pairs of DNA-reading heads (residues 12–55) for interactions with tandem P'1–P'2 arm sites (Fig. 2b–d). The convoluted packing

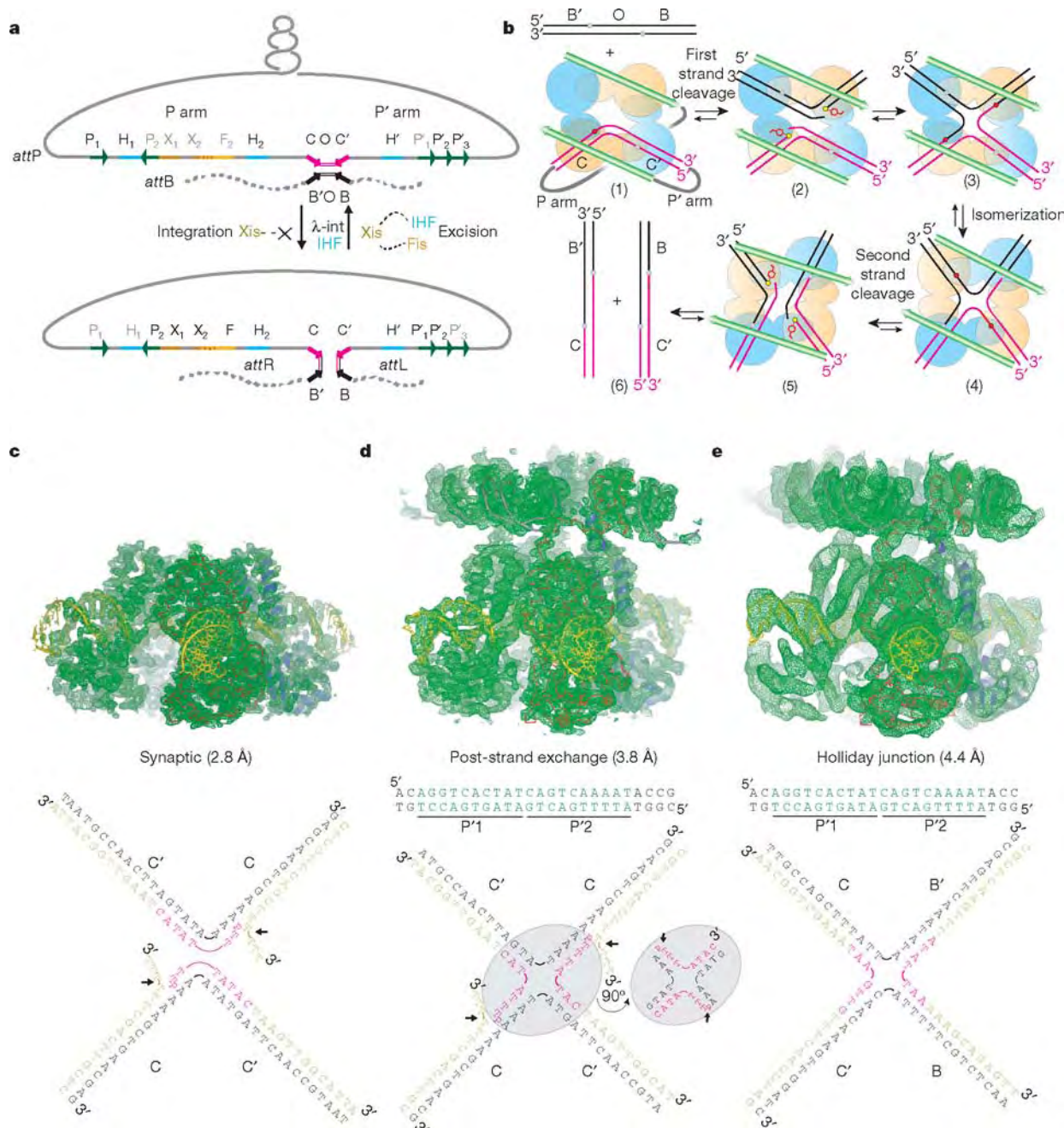


Figure 1 | Pathways for integrative and excisive recombination. **a**, λ -int recombines bacteriophage *attP* and bacterial *attB* attachment sites during integration, and *attL* and *attR* during excision. Different sets of DNA bending proteins (IHF, Xis and Fis) uniquely configure the recombinogenic complexes for these reactions and facilitate simultaneous interaction of the heterobivalent λ -int with the core and arm sites¹ (see Fig. 5). The occupied (black labels) and unoccupied (grey labels) sites during each recombination

process are shown for λ -int, IHF (cyan), Xis (brown) and Fis (orange). **b**, Conformation of the core DNAs and interactions of the N domains (small ellipsoids) with the arm DNAs (green arrows) during a recombination reaction. **c–e**, Experimental electron densities of λ -int complexes after phase modification (top). DNA substrates used for crystallization (see Fig. 3) along with the alternative isomer of the core DNA shown in the inset of **d** (bottom; see Methods).

arrangement of the four N domains buries 2,900 Å² of protein surface area, suggestive of a stable structure once the arms are engaged²⁰. The structure of the N-domain layer and its interactions with arm DNAs are very similar in both post-exchange and Holliday junction complexes of λ -int despite different crystal-packing environments (Supplementary Table S1), a further indication of a fairly rigid structure. The two-fold symmetry of the assembled N-domain layer is reflected to varying extents in the skewed arrangement of catalytic domains and associated core DNAs (Figs 2b and 3d, g). Our results suggest that interactions with the arm DNAs bias the conformation of the λ -int tetramer to drive the reaction to completion.

Four N domains are embraced by two anti-parallel arm DNAs that bend towards one another (13° bend), although not as strongly as inferred from solution studies^{11,23}. Pairs of N domains engage adjacent binding sites (5'-GTCA-3') that are separated by 10 bp (Fig. 2b, c). Each reading head is positioned in the major groove by interactions with the phosphosugar backbone involving side chains on the ends of strand β 1 (Asn 15, Asn 20) and main chain nitrogens from strand β 3 (Glu 34, Gly 36; Supplementary Fig. S1). Side chains

from β 1 (Tyr 17) and the C-terminal end of β 2 (Arg 27) make sequence-specific polar interactions with the base pairs of the P'1 and P'2 arm sites. A basic N-terminal segment (residues 2–10) that is required for recombination activity²⁴ dips into the minor groove adjacent to the 3' side of the consensus arm binding sequence. Patchy electron density indicates that this segment is not well ordered in the presence of arm DNA, and it was entirely disordered in an NMR structure of the unliganded N domain²⁴.

The CB and catalytic domains of λ -int together constitute a minimal recombinase that is structurally analogous to the full-length Cre¹⁵, Flp¹⁷ and XerC/D²⁵ tyrosine recombinases²¹. The active site of λ -int resembles that of other tyrosine recombinases^{19,26}, with conserved residues (Arg 212, Lys 235, His 308, Arg 311, His 333) that engage the scissile phosphate and a tyrosine nucleophile (Tyr 342) that forms a 3' phosphotyrosine linkage with the cleaved DNA¹⁴. The catalytic domain of λ -int binds to core sites with low affinity¹⁰ and has significant DNA cleavage activity only when linked to the CB domain, which contributes binding affinity and specificity for core sites²¹.

Although the CB domain of λ -int resembles the N-terminal

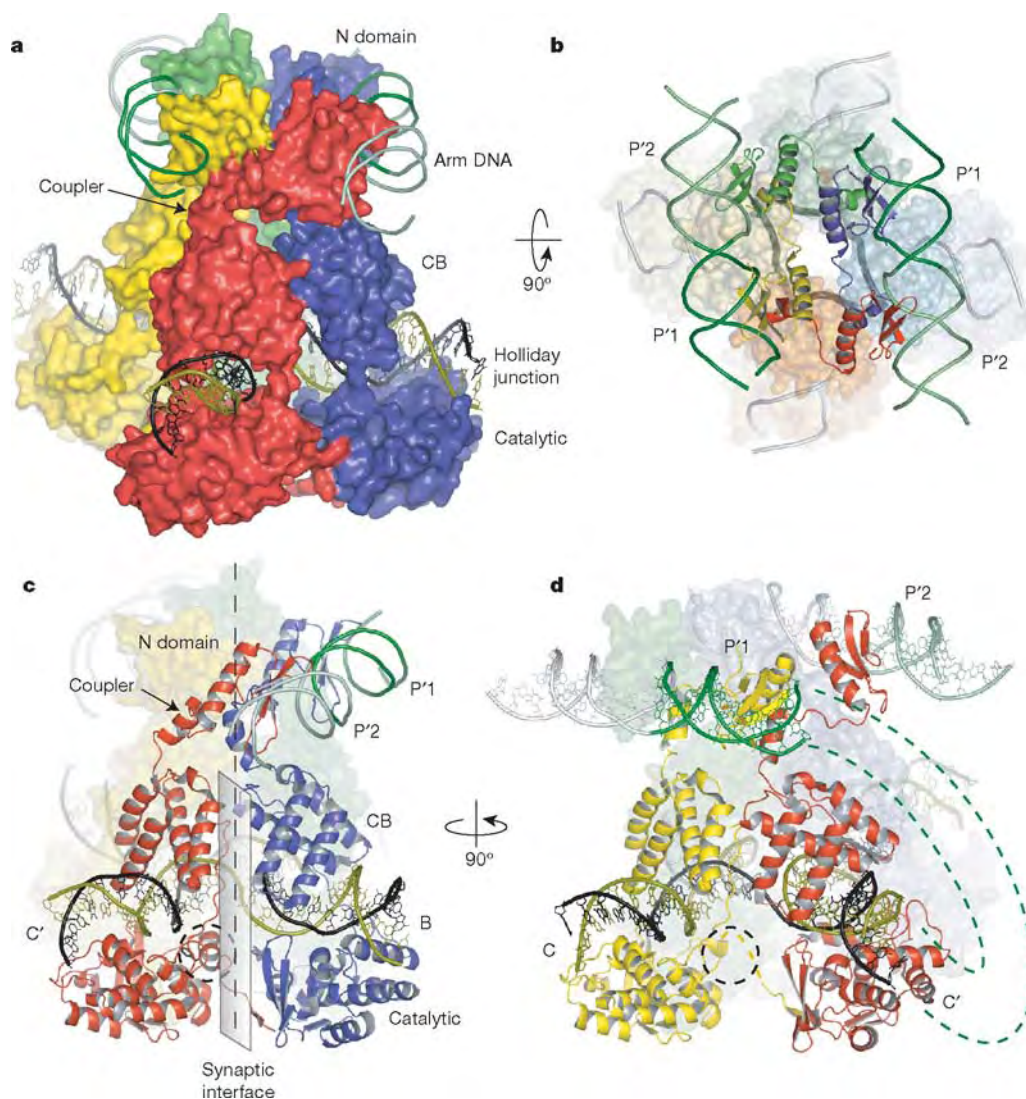


Figure 2 | Structure of the λ -int tetramer bound to a Holliday junction and arm DNAs. **a**, The domains of λ -int pack together as three stacked layers, with the N domains cyclically swapped onto neighbouring subunits. The N-domain layer embraced by two anti-parallel arm DNAs is linked through short α -helical couplers to the CB and catalytic domains, which encircle the branches of the Holliday junction. The active subunits are coloured

red/green and inactive subunits are blue/yellow. **b**, The two-fold symmetry of the N-domain layer is reflected in the skewed arrangement of CB domains and the shape of the four-way junction (thick dark grey lines) in the bottom strands reactive isomer. **c**, **d**, Active (**c**) and inactive (**d**) interfaces are shown, with the area around Tyr 342 nucleophile marked by dashed circles (see Fig. 4).

domain of Cre, it lacks a helix (α E) that contributes many inter-subunit interactions in the Cre tetramer¹⁵. Without this helix, the loosely packed CB domains of the λ -int tetramer rotate against one another by as much as 30° in the different isomers that were crystallized (Fig. 3). The N domains do not rotate and the α -helical coupler connecting the N-domain layer to the CB domains (Fig. 2a, c) appears to restrict the range of subunit rotation. The coupler is essential for cooperative binding to arm sites^{20,27}, and it is well positioned to transmit the allosteric effects of arm binding to the catalytic core.

Isomerization and control of catalytic activity

The complex between λ -int and its DNA substrates isomerizes halfway through the recombination reaction, redirecting DNA cleavage activity from one pair of strands to the other pair for resolution of the Holliday junction (Fig. 1b). The three independent crystal structures represent different steps of isomerization: a synaptic complex (Figs 1c and 3a, b), a post-strand exchange complex

(Figs 1d and 3d, e) and a Holliday junction complex (Figs 1e and 3g, h). Each step reveals a different conformation of the core DNAs and different subunit packing interactions. Conformational transitions between steps require three interrelated modes of motion: a migration of the DNA bend that generates different base-pairing combinations at the centre of the crossover region of the four-way junction, a scissoring motion of the DNA branches extending from the junction, and a sliding of the two halves of the complex past one another that skews its shape away from four-fold symmetry (Fig. 3). These snapshots of the reaction pathway suggest how DNA cleavage is regulated and reveal the molecular features promoting strand exchange.

λ -int's catalytic activity is controlled by the quaternary structure of the tetramer, resulting in two active and two inactive subunits that switch roles during isomerization (Fig. 1b). The experimentally phased electron density, even at the limited resolution of these structures (Fig. 1c–e; see also Supplementary Table S1), shows clear evidence of structural differences among the subunits of the λ -int

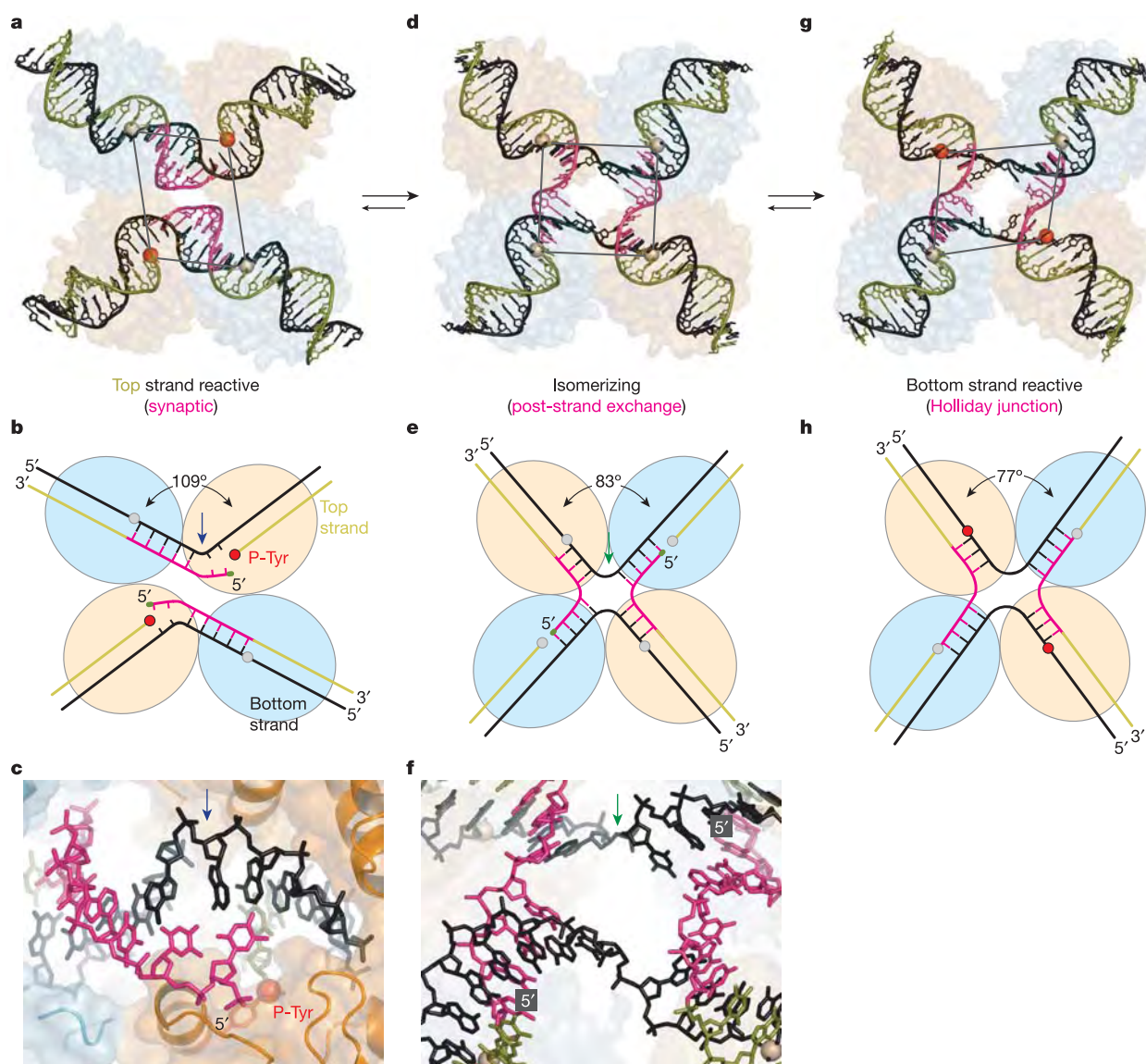


Figure 3 | Three different conformations of λ -int tetramers representing distinct steps of the recombination reaction. The core DNAs within the λ -int(75–356) synaptic complex (a, b), the λ -int post-strand exchange complex (d, e) and the λ -int Holliday junction complex (g, h) are shown along with schematic diagrams illustrating the inter-branch angles and position of branch points. The pair of λ -int subunits in the active

conformation (orange) is positioned closer to the centre of each complex, whereas the inactive pair of subunits (cyan) is further apart. Scissile phosphates (spheres) activated for cleavage are coloured in red. A sharp kink in the bottom strand located two (c; synaptic) or four (f; post-exchange) nucleotides from the site of top strand cleavage is marked with arrows.

tetramer, corresponding to active (red subunit in Fig. 2c) and inactive (yellow subunit in Fig. 2d) conformations of the helix containing the Tyr 342 nucleophile. These two states are most clearly illustrated by the 2.8 Å structure of the synaptic complex (Fig. 4). The catalytic domains interact through a C-terminal extended segment (residues 350–356, including $\beta 9$) that packs *in trans* against the neighbouring subunits (Figs 2c, d and 4a) in a cyclically permuted arrangement similar to that of the N domains (Fig. 2a). The *trans* interactions of λ -int's catalytic domains are functionally analogous to the *trans* packing of active site control elements of Cre and FLP recombinases^{13,14}. The skewed packing of subunits gives rise to two very different subunit interfaces harbouring active or inactive catalytic sites (Figs 2c, d and 4). In the active conformation the Tyr 342 helix is well ordered, stabilized by electrostatic interactions involving the catalytically essential residues Arg 346 and Arg 348 (ref. 28) (Fig. 4b). The position of $\beta 9$ in the inactive subunit is incompatible with these stabilizing interactions and the region around Tyr 342 is correspondingly disordered (Fig. 4c). The active and inactive subunit interfaces (Fig. 2d) also differ by a significant rotation (up to 37°) of the apposing subunits.

The α -helical conformation around Tyr 342 differs from the active site conformation previously reported for a crystal structure of a λ -int(75–356) monomer using a similar flap-type DNA substrate with a phosphorylated 5' end¹⁹ (Fig. 1d). To remove the influence of the added 5' phosphate we determined an additional structure of λ -int(75–356) complexed to a nicked COC' DNA with an unmodified 5'-OH end and orthovanadate bridging between Tyr 342 and the 3'-OH of the nicked DNA^{29,30} (Supplementary Table S1). The active subunits of the vanadate complex adopt the α -helical conformation around Tyr 342 (not shown), suggesting that this is the true active conformation.

DNA bending and strand exchange

The synaptic complex of λ -int(75–356) without arm DNAs represents an early stage of recombination with the 5' ends of the cleaved DNA strands unexchanged (Fig. 3a, b). This synaptic complex deviates strongly from four-fold symmetry, with the scissile phosphates (represented as the corners of a parallelogram) of the cleaved strands located 39 Å apart and those of the uncleaved strands 50 Å apart (Fig. 3a, b). This skewed arrangement of subunits and cleavage sites produces a translational offset between the synapsed DNAs that brings the cleaved 5' ends closer to the 3' phosphotyrosine of the synapsing partner to facilitate strand exchange and ligation. The core DNAs are bent asymmetrically in the overlap region between the sites of DNA cleavage (Fig. 3b, c). A kink in the uncleaved bottom strand located two base pairs away from the cleaved site causes an

unstacking of two adjacent adenines, generating a 109° angle between the C and C' binding sites. A break in the DNA backbone facilitates the localized kinking of DNA that disrupts base pairing interactions with two nucleotides at the 5' end of the cleaved top strand (Fig. 3c). Thus, the sharp kinking of the DNA substrate may facilitate strand exchange and prevent religation.

In the complex with arm DNAs, the cleaved 5' ends of the core DNAs have been exchanged and the core DNAs resemble a Holliday junction with approximate four-fold symmetry (Fig. 3d, e). Ligation of the exchanged strands is prevented by an added 5' phosphate¹⁹, and this may have prevented further isomerization of the complex to a conformer that is proficient for the second strand exchange reaction. In this post-exchange complex, the kink in the overlap region has shifted to a more central location 4 bp away from the cleaved site (Fig. 3d, e). The angle between the C and C' binding sites has also changed (Fig. 3b, e), switching base pairing partners to promote exchange of the 5' ends. This remodelling occurs by local adjustments to the DNA backbone without requiring a significant rotation of the DNA branches radiating from the centre of the four-way junction. The central position of the DNA bend in the post-exchange complex brings the cleavage sites on the bottom strands closer together (44 Å), and this may help prevent the reversal of top strand cleavage. Strand exchange appears to be promoted by a reshaping of the DNA substrate brought about by the binding of the arm DNAs (Figs 1b and 4). There are few interactions with the 5' ends of the cleaved DNA to provide a more directed mechanism.

The recombination complex with a Holliday junction and arm DNAs (Fig. 3g, h) could, in principle, be resolved by the exchange of either pair of DNA strands, and this choice is strongly influenced by the position of the crossover junction within the overlap region^{18,31}. The Holliday junction used for crystallization (Fig. 1e) is immobile, with its crossover point fixed three nucleotides from one pair of cleavage sites and, consequently, these sites are used preferentially for resolution¹⁸. The complex is highly skewed, and the scissile phosphates bound by active subunits are brought close together (38 Å) after the initial pair of strands have exchanged (Fig. 3g, h). For reasons described below, we propose that this complex represents an isomer poised for the second strand exchange that would resolve the Holliday junction into recombinant products.

Because recombination involves two analogous strand exchange reactions¹⁸ (Fig. 1b), we could reasonably expect that complexes catalysing the first and second strand exchanges will have similar conformations and be distinguished only by the identities of the exchanging strands. The arm DNAs bind in a unique orientation (Fig. 2c, d; see Methods). As a result, the missing connections with the core sites can be readily modelled to establish which core sites are

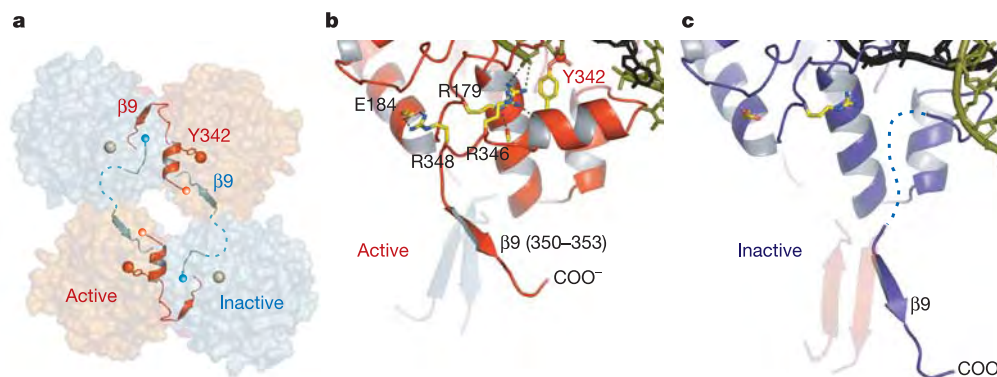


Figure 4 | Control of catalytic activity by *trans* cyclic interactions of the C-terminal tail. **a**, The assembly of active (red) and inactive (cyan) catalytic sites results from a skewed packing arrangement of λ -int(75–356) subunits in the tetramer. The scissile phosphates bound by active and inactive subunits are shown as red and grey spheres, respectively. **b**, In the active

conformation, the tail is oriented to permit stabilizing interactions with residues Arg 346 and Arg 348 that result in a well-ordered segment (residues 339–348) containing the Tyr 342 nucleophile. **c**, In the inactive molecule, a large shift of $\beta 9$ causes this segment to become disordered.

bound by active and inactive subunits (Fig. 5; see also Supplementary Information). On the basis of the proximity to the P'1 arm sites, we propose that the C' sites (sites of bottom strand cleavage)^{3,4} are bound by subunits in the active conformation in both complexes with arms (Fig. 2d). Correspondingly, biochemical assays have shown that interactions with arm DNAs stimulate Holliday junction resolution by bottom strand exchange²².

Although the positions of active and inactive subunits are defined with respect to the arms, the COC' core DNAs freely isomerize, and X-ray experiments with iodinated DNAs reveal two isomers of the cleaved COC' DNAs bound in different orientations (Fig. 1d). The majority of the DNA molecules have isomerized so that the cleaved DNA strands are bound by subunits now in an inactive conformation. The existence of both isomers in the same crystal highlights the reciprocal nature of the isomerization process when the core and arm-binding sites are supplied on separate DNAs. Nonetheless, the unique binding orientation of the arm DNAs specifies a topological linkage to the core sites of bottom strand cleavage, which are bound by subunits in the active conformation.

Allosteric control of recombination

How do interactions with the arms promote the exchange of bottom strands that completes λ integration and excision^{3,4} (Fig. 1b, step 5)? The similarly skewed geometry of the N- and C-terminal halves of λ recombination complexes raises the possibility that arm-binding interactions may restrain the quaternary structure of the λ -int tetramer. The N domains in complex with arm sites are arranged in a parallelogram (Fig. 2b), mirroring the skewed arrangement of scissile phosphates configured for bottom strand exchange (Fig. 3g). One pair of N domains lies near the two-fold axis of symmetry and the other pair lies further away, reflecting the arrangement of active and inactive subunits. It appears that the α -helical coupler joining the N domain with the CB domain (Fig. 2c) functions as a short

tether that imposes the skewed, two-fold symmetry of the N-domain layer onto the four CB domains and associated core DNAs.

This is a type of allosteric control that positions the bottom strands for cleavage and exchange (Fig. 1b), effectively shifting the reaction equilibrium in favour of recombinant products. Despite the conformational bias favouring bottom strand exchange, the exchange of top strands is a prerequisite that removes a physical barrier from the path of bottom strand exchange¹³. It is therefore notable that arm binding also establishes an orientation of core DNAs that is compatible with top strand cleavage in the initial synaptic complex. Cleavage of the top strands may render the DNAs more flexible, allowing an inter-conversion to the bottom strands reactive isomer that is stabilized by the N-domain layer.

Interactions with the arms of attP

Although the basic features of the DNA cleavage and strand exchange reactions in the λ recombination system are similar to those of other tyrosine recombinases^{13,14,32}, the arms of the bacteriophage λ att sites provide a level of regulation and unidirectionality that is not available to simpler recombinases³³. The decision to insert bacteriophage λ DNA into the *Escherichia coli* chromosome or to excise it is made by committees of DNA bending proteins that wrap the P and P' arms around the integrative and excisive complexes in different ways^{1,34,35}. Our crystal structures of λ -int tetramers reveal the spatial relationship between arm and core sites that together with the crystal structure of IHF bound to DNA³⁶ can be used to model the topology of the arms in both complexes (Fig. 5).

The conformations of the P' arm are readily explained by simple docking models that are similar for both integrative and excisive complexes (Figs 1 and 5a). IHF binding to the H' site (Fig. 1a) creates an approximately 180° bend that redirects the P' arm over the λ -int tetramer for interactions with two N domains located on opposing sides of the synaptic interface^{36,37}. Only a minor change in the

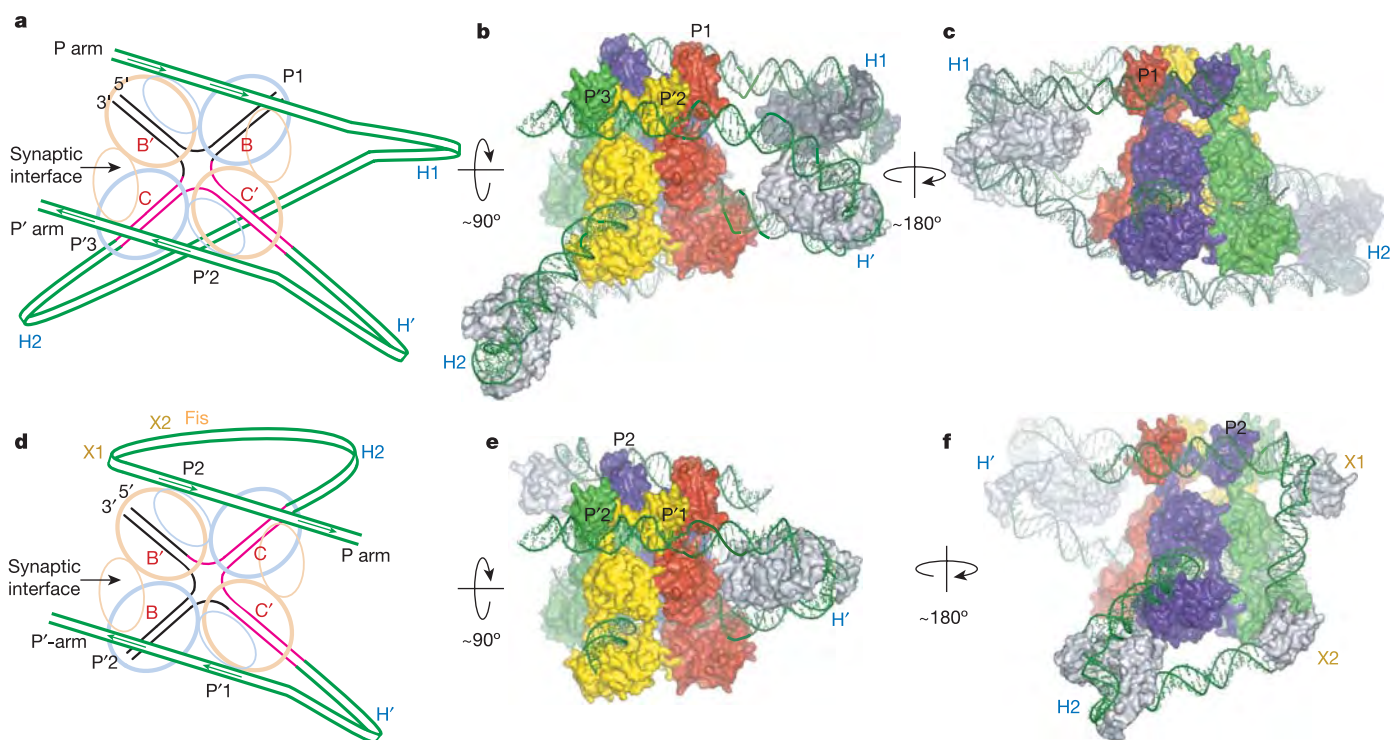


Figure 5 | Models showing the topology of DNA during λ integration and excision. **a**, A schematic representation of the integrative complex. The P' arm engages the N domains of subunits bound to C and B' core sites. The P arm is positioned over the N domains of subunits on the C' and B core sites by a compound bend induced by IHF. **d**, During excision, the P' arm is

oriented similarly as in integration, whereas the P arm could form a more compact loop upon binding to Xis and/or Fis (see text). **b**, **c**, **e**, **f**, Docking models of the hypothetical integrative (**b**, **c**) and excisive (**e**, **f**) complexes based on our crystal structures and that of an IHF–DNA complex³⁶ are shown in front and back views.

bending angle around H' is required to alternatively engage P'2–P'3 (integration) or P'1–P'2 (excision) binding sites⁸ (Fig. 5b, e).

The P arm of the phage attachment site is configured differently during integration and excision^{8,38}, engaging either the P1 or P2 site, respectively (Fig. 1a). Different pairs of attachment sites are used for integration (*attP* and *attB*) or excision (*attL* and *attR*) with resultant differences in the relative orientations of P and P' arms (Fig. 5a, d). Our modelling suggests that during integration IHF bends the P arm at H2 and directs it under the λ -int tetramer. A second IHF-induced bend at the distal H1 site would position the P1 site over one N domain, and the adjacent N domain binds nonspecifically to the flanking DNA (Fig. 5a–c). During excision, IHF is displaced from H1 by the DNA-bending factors Xis and Fis that bind nearby sites X1 and X2 or F^{1,6}. Presumably, a sharp bend in the P arm, induced by Xis²³ binding to DNA and its interaction with the N domain³⁹, is required to bring the P2 site into position. We have modelled one such conformation of P arm in the excisive complex (Fig. 5d–f). In both reactions, we predict that the P arm wraps around the complex with a left-handed curvature that would be stabilized in underwound DNA. Negative supercoiling of *attP* is required for integration⁴⁰, and this may compensate for the weaker interactions of the P arm with λ -int.

The crystal structures of λ -int tetramers and the topology of arm binding inferred from them indicate how the tetramer may initially assemble by the coalescence of N domains in complex with high-affinity arm sites (Fig. 2) to enable the low-affinity interactions with core sites⁴¹. During integration, the interlocked assembly of N domains and arm DNAs positions two subunits on the *attP* core and establishes a preferred bending direction for top strand cleavage. Two remaining subunits of the tetramer are then available for the capture of *attB*, consistent with earlier DNA footprinting studies⁴². During excision, the arm DNAs create a bridge between subunits across the synaptic interface of *attL* and *attR*, dictating a bend in the core DNA that again initially favours top strand cleavage before isomerization and exchange of the bottom strands. Arm binding interactions appear to bias isomerization in favour of the second half of the reaction: the exchange of bottom strands and resolution into products (Figs 1b and 3). The structures of higher-order λ recombinogenic complexes show how the simultaneous binding of λ -int to core and arm sites facilitates synapsis, and they suggest a mechanism for coordinating the order of strand exchange and driving the multi-step recombination reaction to completion.

METHODS

The preparation of λ -int proteins and the DNA complexes that were crystallized are described in the Supplementary Information. X-ray diffraction data were collected at beamlines X-25 and X-26C of National Synchrotron Light Source, and beamline 19-ID of Advanced Photon Source. X-ray data were processed with HKL2000 (ref. 43) and CCP4⁴⁴ program suites. The structure of the λ -int(75–356) complex (synaptic) was determined by the single-wavelength anomalous dispersion (SAD) method using a selenomethionine (SeMet)-labelled crystal. The structure of the λ -int–Holliday junction–arm DNA complex was determined by a combination of multiple-wavelength anomalous dispersion (MAD) and conventional isomorphous replacement phasing, using SeMet, Ta₆Br₁₄·8H₂O and Pt-terpyridine derivatives. Heavy atom derivatives were prepared by soaking crystals for 12 h at 20 °C in mother liquor containing 0.4 mM tantalum bromide or 0.1 mM platinum terpyridine. The structure of the λ -int–COC'–arm DNA complex (post-strand exchange) was determined with phase information from SeMet and Ta₆Br₁₄·8H₂O derivatives. MLPHARE⁴⁴ and SHARP⁴⁵ were used for heavy-atom parameter refinement and calculation of experimental phases, which were subsequently improved by density modification using DM⁴⁴, RESOLVE⁴⁶ and SOLOMON⁴⁴ (Supplementary Table S1). Atomic models for all structures were fitted into the electron density using the program O⁴⁷, starting with the refined model of a λ -int(75–356) DNA complex and the NMR solution structure²⁴ of the N domain. The register and orientation of the core and arm DNAs in the λ -int structures were determined from X-ray experiments using DNA substrates containing 5'-iodo-2'-deoxyuridine substitutions at multiple thymidine positions, and the locations of covalent 3' phosphotyrosine linkages. The λ -int(75–356) structure was refined using

CNS⁴⁸ without restraints of non-crystallographic symmetry (NCS) to values of $R_{\text{work}}/R_{\text{free}} = 0.218/0.262$ at 2.8 Å. The λ -int–COC'–arm structure was initially refined using CNS with a grouped B-factor refinement, followed by REFMAC5⁴⁴ with positional and TLS parameter refinement to a final $R_{\text{work}}/R_{\text{free}} = 0.248/0.296$ at 3.8 Å. The λ -int–Holliday junction–arm structure was refined using REFMAC5 with the rigid body, TLS and positional refinements, resulting in $R_{\text{work}}/R_{\text{free}} = 0.281/0.316$ at 4.4 Å. Experimental phases were included as refinement standards in all cases. NCS restraints were imposed on the individual domains during refinement of the λ -int–core–arm structures, and hydrogen bonding restraints were used for protein secondary structures and DNA base pairs. The figures were produced with PyMOL (<http://www.pymol.org>).

Received 3 December 2004; accepted 15 April 2005.

1. Landy, A. Dynamic, structural, and regulatory aspects of λ site-specific recombination. *Annu. Rev. Biochem.* **58**, 913–949 (1989).
2. Azaro, M. A. & Landy, A. in *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A.) 118–148 (ASM Press, Washington DC, 2002).
3. Kitts, P. A. & Nash, H. A. Bacteriophage lambda site-specific recombination proceeds with a defined order of strand exchanges. *J. Mol. Biol.* **204**, 95–107 (1988).
4. Nunes-Duby, S. E., Matsumoto, L. & Landy, A. Site-specific recombination intermediates trapped with suicide substrates. *Cell* **50**, 779–788 (1987).
5. Segall, A. M. & Nash, H. A. Architectural flexibility in lambda site-specific recombination: three alternate conformations channel the attL site into three distinct pathways. *Genes Cells* **1**, 453–463 (1996).
6. Nash, H. A., et al. *Escherichia coli* and *Salmonella typhimurium*: in *Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 2263–2376 (ASM Press, Washington DC, 1996).
7. Moitoso de Vargas, L., Kim, S. & Landy, A. DNA looping generated by DNA bending protein IHF and the two domains of lambda integrase. *Science* **244**, 1457–1461 (1989).
8. Numrych, T. E., Gumport, R. I. & Gardner, J. F. A comparison of the effects of single-base and triple-base changes in the integrase arm-type binding sites on the site-specific recombination of bacteriophage lambda. *Nucleic Acids Res.* **18**, 3953–3959 (1990).
9. Ross, W. & Landy, A. Bacteriophage λ int protein recognizes two classes of sequence in the phage att site: characterization of arm-type sites. *Proc. Natl Acad. Sci. USA* **79**, 7724–7728 (1982).
10. Ross, W., Landy, A., Kikuchi, Y. & Nash, H. Interaction of int protein with specific sites on λ att DNA. *Cell* **18**, 297–307 (1979).
11. Radman-Livaja, M. et al. Arm sequences contribute to the architecture and catalytic function of a λ integrase-Holliday junction complex. *Mol. Cell* **11**, 783–794 (2003).
12. Franz, B. & Landy, A. The Holliday junction intermediates of λ integrative and excisive recombination respond differently to the bending proteins integration host factor and excisionase. *EMBO J.* **14**, 397–406 (1995).
13. Van Duyne, G. D. A structural view of cre-loxP site-specific recombination. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 87–104 (2001).
14. Chen, Y. & Rice, P. A. New insight into site-specific recombination from Flp recombinase-DNA structures. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 135–159 (2003).
15. Guo, F., Gopaul, D. N. & van Duyne, G. D. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* **389**, 40–46 (1997).
16. Gopaul, D. N., Guo, F. & Van Duyne, G. D. Structure of the Holliday junction intermediate in Cre-loxP site-specific recombination. *EMBO J.* **17**, 4175–4187 (1998).
17. Chen, Y., Narendra, U., Iype, L. E., Cox, M. M. & Rice, P. A. Crystal structure of a Flp recombinase-Holliday junction complex: assembly of an active oligomer by helix swapping. *Mol. Cell* **6**, 885–897 (2000).
18. Azaro, M. A. & Landy, A. The isomeric preference of Holliday junctions influences resolution bias by λ integrase. *EMBO J.* **16**, 3744–3755 (1997).
19. Aihara, H., Kwon, H. J., Nunes-Duby, S. E., Landy, A. & Ellenberger, T. A conformational switch controls the DNA cleavage activity of λ integrase. *Mol. Cell* **12**, 187–198 (2003).
20. Sarkar, D. et al. Differential affinity and cooperativity functions of the amino-terminal 70 residues of λ integrase. *J. Mol. Biol.* **324**, 775–789 (2002).
21. Tirumalai, R. S., Kwon, H. J., Cardente, E. H., Ellenberger, T. & Landy, A. Recognition of core-type DNA sites by λ integrase. *J. Mol. Biol.* **279**, 513–527 (1998).
22. Radman-Livaja, M., Biswas, T., Mierke, D. & Landy, A. Architecture of recombination intermediates visualized by in-gel FRET of λ integrase-Holliday junction-arm-DNA complexes. *Proc. Natl Acad. Sci. USA* **102**, 3913–3920 (2005).
23. Thompson, J. F. & Landy, A. Empirical estimation of protein-induced DNA bending angles: applications to lambda site-specific recombination complexes. *Nucleic Acids Res.* **16**, 9687–9705 (1988).
24. Wojciak, J. M., Sarkar, D., Landy, A. & Clubb, R. T. Arm-site binding by λ -integrase: solution structure and functional characterization of its amino-terminal domain. *Proc. Natl Acad. Sci. USA* **99**, 3434–3439 (2002).

25. Subramanya, H. S. *et al.* Crystal structure of the site-specific recombinase, XerD. *EMBO J.* **16**, 5178–5187 (1997).
26. Kwon, H. J., Tirumalai, R., Landy, A. & Ellenberger, T. Flexibility in DNA recombination: structure of the lambda integrase catalytic core. *Science* **276**, 126–131 (1997).
27. Warren, D., Lee, S. Y. & Landy, A. Mutations in the amino-terminal domain of λ -integrase have differential effects on integrative and excisive recombination. *Mol. Microbiol.* **55**, 1104–1112 (2005).
28. Kazmierczak, R. A., Swalla, B. M., Burgin, A. B., Gumpert, R. I. & Gardner, J. F. Regulation of site-specific recombination by the C-terminus of lambda integrase. *Nucleic Acids Res.* **30**, 5193–5204 (2002).
29. Davies, D. R., Interthal, H., Champoux, J. J. & Hol, W. G. Crystal structure of a transition state mimic for Tdp1 assembled from vanadate, DNA, and a topoisomerase I-derived peptide. *Chem. Biol.* **10**, 139–147 (2003).
30. Martin, S. S., Wachi, S. & Baldwin, E. P. Vanadate-based transition-state analog inhibitors of Cre-LoxP recombination. *Biochem. Biophys. Res. Commun.* **308**, 529–534 (2003).
31. Nunes-Duby, S. E., Azaro, M. A. & Landy, A. Swapping DNA strands and sensing homology without branch migration in λ site-specific recombination. *Curr. Biol.* **5**, 139–148 (1995).
32. Sherratt, D. J. & Wigley, D. B. Conserved themes but novel activities in recombinases and topoisomerases. *Cell* **93**, 149–152 (1998).
33. Bushman, W., Thompson, J. F., Vargas, L. & Landy, A. Control of directionality in lambda site specific recombination. *Science* **230**, 906–911 (1985).
34. Crisona, N. J., Weinberg, R. L., Peter, B. J., Sumners, D. W. & Cozzarelli, N. R. The topological mechanism of phage λ integrase. *J. Mol. Biol.* **289**, 747–775 (1999).
35. Bankhead, T. M. *et al.* Mutations at residues 282, 286, and 293 of phage λ integrase exert pathway-specific effects on synapsis and catalysis in recombination. *J. Bacteriol.* **185**, 2653–2666 (2003).
36. Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87**, 1295–1306 (1996).
37. Swinger, K. K. & Rice, P. A. IHF and HU: flexible architects of bent DNA. *Curr. Opin. Struct. Biol.* **14**, 28–35 (2004).
38. Thompson, J. F., de Vargas, L. M., Skinner, S. E. & Landy, A. Protein-protein interactions in a higher-order structure direct lambda site-specific recombination. *J. Mol. Biol.* **195**, 481–493 (1987).
39. Sam, M. D., Cascio, D., Johnson, R. C. & Clubb, R. T. Crystal structure of the excisionase-DNA complex from bacteriophage lambda. *J. Mol. Biol.* **338**, 229–240 (2004).
40. Mizuuchi, K., Gellert, M. & Nash, H. A. Involvement of supercoiled DNA in integrative recombination of bacteriophage lambda. *J. Mol. Biol.* **121**, 375–392 (1978).
41. Kim, S. & Landy, A. Lambda Int protein bridges between higher order complexes at two distant chromosomal loci attL and attR. *Science* **256**, 198–203 (1992).
42. Richet, E., Abcarian, P. & Nash, H. A. Synapsis of attachment sites during lambda integrative recombination involves capture of a naked DNA by a protein-DNA complex. *Cell* **52**, 9–17 (1988).
43. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
44. Collaborative Computational Project Number 4, The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
45. De LaFortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276**, 472–499 (1997).
46. Terwilliger, T. C. Automated structure solution, density modification and model building. *Acta Crystallogr. D* **58**, 1937–1940 (2002).
47. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
48. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements X-ray data were measured at beamlines X-26C and X-25 of the National Synchrotron Light Source, station A1 of the Cornell High Energy Synchrotron Source, and beamline 19ID at the Advanced Photon Source, facilities that are supported by the US Department of Energy and the National Institutes of Health. This work was supported by research grants from the National Institutes of Health (to T.E. and A.L.) and fellowships from Jeane B. Kempner Fund (to T.B.) and the Human Frontier Science Program (to H.A.). T.E. is the Hsien Wu and Daisy Yen Wu Professor at Harvard Medical School.

Author Information Coordinates and structure factors are deposited in the Protein Data Bank under accession codes 1Z19, 1Z1B and 1Z1G. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.E. (tome@hms.harvard.edu).

A planetary system as the origin of structure in Fomalhaut's dust belt

Paul Kalas¹, James R. Graham¹ & Mark Clampin²

The Sun and >15 per cent of nearby stars are surrounded by dusty disks that must be collisionally replenished by asteroids and comets, as the dust would otherwise be depleted on timescales <10⁷ years (ref. 1). Theoretical studies show that the structure of a dusty disk can be modified by the gravitational influence of planets^{2–4}, but the observational evidence is incomplete, at least in part because maps of the thermal infrared emission from the disks have low linear resolution (35 AU in the best case⁵). Optical images provide higher resolution, but the closest examples (AU Mic and β Pic) are edge-on^{6,7}, preventing the direct measurement of the azimuthal and radial disk structure that is required for fitting theoretical models of planetary perturbations. Here we report the detection of optical light reflected from the dust grains orbiting Fomalhaut (HD 216956). The system is inclined 24° away from edge-on, enabling the measurement of disk structure around its entire circumference, at a linear resolution of 0.5 AU. The dust is distributed in a belt 25 AU wide, with a very sharp inner edge at a radial distance of 133 AU, and we measure an offset of 15 AU between the belt's geometric centre and Fomalhaut. Taken together, the sharp inner edge and offset demonstrate the presence of planetary-mass objects orbiting Fomalhaut.

Fomalhaut's circumstellar dust was discovered at thermal infrared wavelengths with the Infrared Astronomical Satellite^{8,9}, and maps of dust emission at submillimetre wavelengths with 7.5–14" angular resolution suggested a ring of material residing between 100 and 140 AU radius^{10,11}. Fomalhaut is only 7.7 parsecs (1 pc = 3.3 light years) from the Sun. Using the Advanced Camera for Surveys (ACS) on board the Hubble Space Telescope (HST), we detect its dust complex at optical wavelengths, with angular resolution 100 times sharper than previous submillimetre maps (Fig. 1; Methods). Fomalhaut is surrounded by a narrow dust belt inclined to our line of sight and with significant asymmetry in its azimuthal brightness distribution. We perform a nonlinear least-squares fit to the ellipse traced by the brightest points along the belt. The semimajor and semiminor axes are 140.7 ± 1.8 AU and 57.5 ± 0.7 AU, respectively. The position angle of the ellipse is $PA = 156.0^\circ \pm 0.3^\circ$, with inclination to the line of sight $i = 65.9^\circ \pm 0.4^\circ$ (assuming that the structure is intrinsically circular). These values are consistent with the belt properties previously inferred via modelling of submillimetre data¹¹.

The geometric centre of the elliptical fit to Fomalhaut's belt is offset from the star by 13.4 ± 1.0 AU at $PA = 350.4^\circ$. Assuming that the star and belt are coplanar, the projected offset translates to 15.3 AU in the plane of the belt. A keplerian orbit with non-zero eccentricity traces an ellipse where the star is at one focus and the centre of the ellipse is offset by a value, f , equivalent to the semimajor axis, a , times the eccentricity, e . If we assume Fomalhaut's belt is non-circular, then the observed offset and semimajor axis give $e = 0.11 \pm 0.01$. Periastron is at $PA = 170^\circ$ (Fig. 1b). The non-circularity of the structure leads to a small revision for the inclination, $i = 65.6^\circ \pm 0.4^\circ$ (ref. 12).

The offset of the star relative to the dust belt and an asymmetric scattering phase function produce the azimuthal brightness asymmetries. In the rotated frame shown in Fig. 1, we refer to the upper left quadrant of the belt as quadrant one (Q1), with Q2, Q3 and Q4

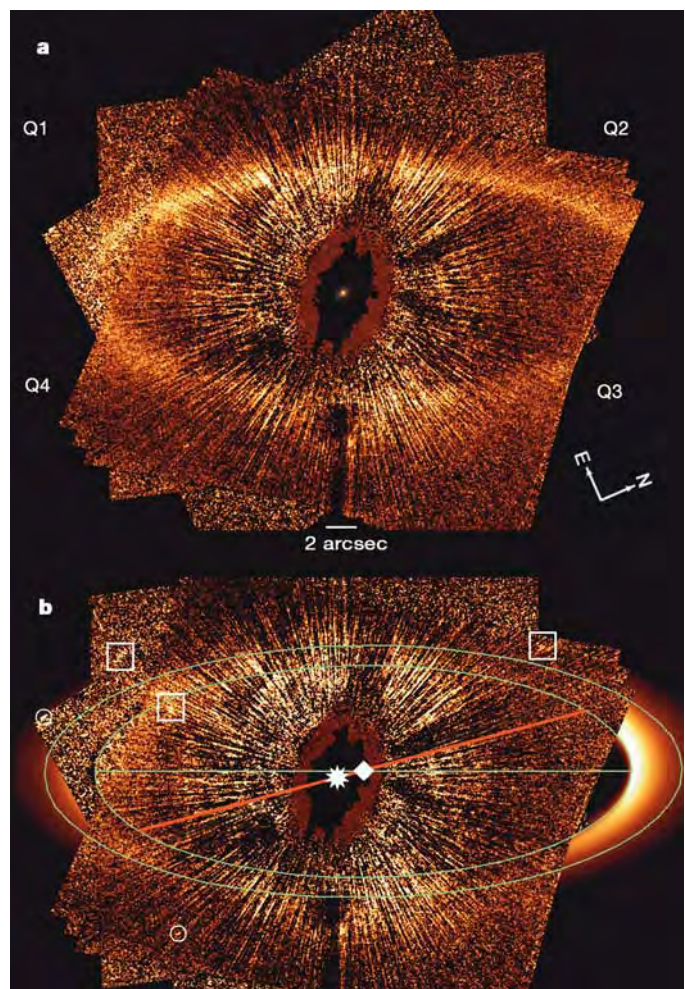


Figure 1 | Optical detection of dust in Fomalhaut's Kuiper belt. **a**, False-colour coronagraphic image (Methods). North is 66° clockwise from vertical. **b**, Residual image after subtracting a model of a dust belt (Methods). Green lines trace the projected boundaries of the detected belt (133–158 AU). The white diamond and asterisk mark the centres of the belt and the star, respectively. The horizontal green line traces the belt semimajor axis, whereas the red line traces the vector between the belt and star centres. White boxes and circles mark extended objects and background stars, respectively.

¹Astronomy Department, University of California, Berkeley, California 94720, USA. ²Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.

Table 1 | Properties of trans-planetary belts

	Fomalhaut	Sun*
Age (Myr)	200 ± 100†	4,600
Spectral type	A3V	G2V
Stellar mass (solar mass)	2.0	1.0
Semimajor axis (AU)	141	42, 48, 39‡
Inner radius (AU)	133	38, 25, 41‡
Outer radius (AU)	158	48, 53, 55‡
Belt dust mass (g)	1.1 × 10 ²⁶ §	6.0 × 10 ²²
Albedo	0.05–0.1	0.07
Outermost planet (AU)		30
Symmetry offset (AU)	15.3	0.01
Mean eccentricity	0.11	0.09, 0.14, 0.36‡
Mean inclination (degrees)	1.5	7.1, 9.8, 12.9‡

* Kuiper belt properties from ref. 27, except for belt dust mass from ref. 23.

† Ref. 24.

‡ First, second and third entries are the mean values for classical Kuiper belt, the 2:1 resonance population, and the 3:2 resonance population, respectively.

§ Ref. 10.

|| Value for the zodiacal dust cloud from ref. 28.

clockwise from Q1. We sample the surface brightness in these quadrants at angles 25° away from the horizontal axis, giving 20.6 ± 0.1 , 21.0 ± 0.1 , 21.5 ± 0.1 and 21.3 ± 0.3 mag arcsec⁻² for Q1, Q2, Q3 and Q4, respectively. The absolute photometry is uncertain by ~ 0.4 mag arcsec⁻² owing to errors in determining the global sky background. However, the azimuthal variation in relative surface brightness indicated above is a consistent finding in different data sets (Methods). Because the star is closer to the belt in Q1 and Q4, these quadrants appear brighter than the corresponding regions in Q2 and Q3. The fact that half the belt, Q1 + Q2, is brighter than the other half, Q3 + Q4, points to an asymmetric scattering phase function. To search for azimuthal asymmetry in the belt that does not arise from the centre of symmetry offset, we deproject the data and multiply each element in the belt by the square of its radius from the star (Fig. 2). The phase function asymmetry is still evident, but with no evidence for brightness asymmetry between Q1 and Q2, or between Q3 and Q4 (Fig. 2b).

We produce a scattered light model of an optically thin dust belt where we vary the belt geometry and the asymmetry parameter, g , in a Henyey–Greenstein phase function (Methods; ref. 13). We iteratively subtract models from the data until Fomalhaut's belt disappears from

the data (Fig. 1b), arriving at model parameters $|g| = 0.2$ and the belt inner edge at 133 AU. The magnitude of the asymmetry parameter is consistent with the observed phase function of zodiacal light particles ($g \approx +0.2$; forward scattering), and with other debris disks (for example, HD 141569; ref. 14). Particles in the zodiacal dust cloud have a median size of $\sim 30 \mu\text{m}$ (ref. 15), which is consistent with the $\sim 100 \mu\text{m}$ grain sizes inferred for Fomalhaut based on model fits to thermal infrared data^{11,16}. Moreover, grains smaller than $\sim 7 \mu\text{m}$ are ejected from the Fomalhaut system by radiation pressure^{17,18}. If the grains surrounding Fomalhaut are in fact comparable to our interplanetary dust grains, then the forward scattering direction (the side pointing out of the sky plane) is the top half of the belt (Q1 + Q2).

The integrated light from the model belt is 16.2 mag, which corresponds to a total grain scattering cross-section $C = 5.5 \times 10^{25} \text{ cm}^2/A$, where A is the grain albedo. Given a steady state collisional size distribution $dN = Ka^{-3.5} da$, where K is a constant and N is the number of grains with radius a , the total disk mass, M_D , is

$$M_D = \frac{4C\rho[a_{\text{max}}^{1/2} - a_{\text{min}}^{1/2}]}{3[a_{\text{min}}^{-1/2} - a_{\text{max}}^{-1/2}]} \quad (1)$$

where a_{min} and a_{max} are the minimum and maximum grain sizes, respectively, and ρ is the grain density, assumed to be uniform throughout the belt. Previous submillimetre observations at 450–850 μm are most sensitive to grain sizes 0.1–1.5 mm, and the Fomalhaut submillimetre data give $M_D = 1.1 \times 10^{26} \text{ g}$ (Table 1). Inserting $a_{\text{min}} = 0.1 \text{ mm}$ and $a_{\text{max}} = 1.5 \text{ mm}$ into equation (1), with $\rho = 2.5 \text{ g cm}^{-3}$ and $C = 5.5 \times 10^{25} \text{ cm}^2$ ($A = 1$), we find $M_D = 7.4 \times 10^{24} \text{ g}$. If $A = 0.1$ –0.05, then $M_D = (0.7$ –1.5) $\times 10^{26} \text{ g}$, in agreement with the submillimetre mass. If the optically detected dust and submillimetre detected dust comprise the same grain population, then we conclude that Fomalhaut's belt is composed of relatively dark grains, consistent with Kuiper belt objects and cometary nuclei (Table 1). The total belt mass from $a_{\text{min}} = 7 \mu\text{m}$ to $a_{\text{max}} = 500 \text{ km}$ is $(3.4$ –6.8) $\times 10^{29} \text{ g}$ (50–100 Earth masses) for $A = 0.1$ –0.05, respectively.

The residual image, after model subtraction, emphasizes excess nebulosity interior to the southeast lobe (Q1 and Q4) that intrudes inward from the inner edge of the belt to as close as 100 AU radius

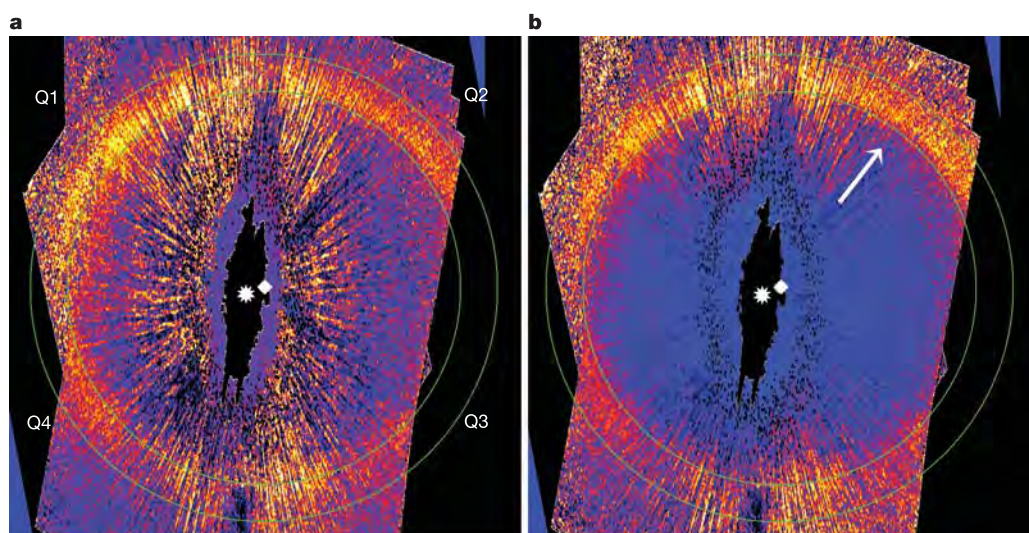


Figure 2 | False colour representations of the deprojected belt image, with symbols as in Fig. 1. a, The observed, de-projected surface brightness. **b**, The surface brightness scaled by distance squared from Fomalhaut, correcting for the inverse square dilution of radiation. The stellar flux normalization is chosen so that it is unity at the centre of the ring (140 AU).

If the Fomalhaut dust belt consists of grains of uniform optical scattering properties, then **b** maps the material surface density for an optically thin distribution. A white arrow indicates the direction of the radial cut in Fig. 3. See text for Q1–Q4.

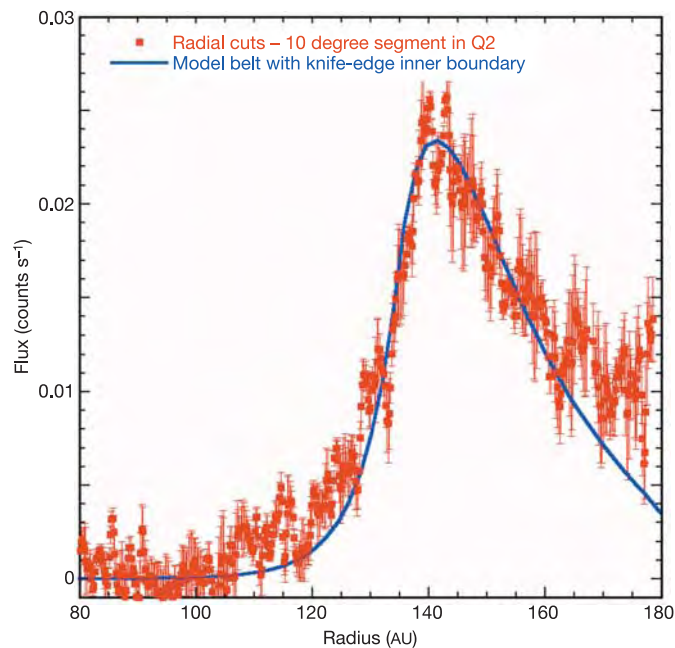


Figure 3 | Radial cut across the belt in the deprojected, illumination-corrected data (Fig. 2). The illumination correction amplifies the sky level and noise at large radii such that the plot shows the sky level beyond ~ 158 AU. The belt was sampled in 2° increments along a total of 10° . The mean values and the standard deviations per measurement are shown. The blue line is an identical radial cut across the model belt (Methods, Fig. 1). The model belt has sharp inner and outer edges, but the observed edges are softened owing to the vertical height of the belt projected onto the sky plane.

from the star along the major axis of the belt (Fig. 1b). No comparable excess is detected interior to the northwest lobe of the belt. Within equivalent regions between 100 and 120 AU radius along the projected semimajor axis on either side of the belt, the southeast interior nebulosity has 1.7 times greater cumulative flux than the northwest side (median surface brightness is 21.6 ± 0.1 and 22.2 ± 0.1 mag arcsec $^{-2}$ for the southeast and northwest, respectively). Fomalhaut's inner dust component was recently detected at thermal infrared wavelengths^{16,19}. We interpret the inward intrusion of nebulosity in the southeast as a tenuous inner dust component that is symmetrical, but detected only in the southeast because it is more strongly illuminated due to the star-belt symmetry offset (Fig. 2b).

Fomalhaut's belt appears narrowest in Q2, near apastron. This result is in contrast to planetary rings, where the narrowest width occurs at periastron owing to rigid precession controlled largely by the self-gravity of ring particles²⁰. Future theoretical work will have to consider that self-gravity in Fomalhaut's belt is probably negligible, given the large distance scales and low mass. A radial cut across the belt in Q2 shows a very sharp inner boundary that is consistent with our model belt (which assumes a knife-edge inner boundary) (Fig. 3). The radial cut shows a plateau at ~ 140 AU that is 1–2 AU wide. The outer boundary of the belt has a less steep fall-off than the inner edge. However, the model adopted to fit the radial drop-off in surface brightness has a particle number density distribution decreasing as r^{-9} , where r is the radius (Methods). This indicates that the disk outer extent also has a steep drop-off, and that the true outer boundary may lie beyond the detection-limited boundary at 158 AU. The model disk adopted here is also very flat, with a 3.5 AU scale height at 133 AU, giving a belt opening angle of $\sim 1.5^\circ$.

The sharp inner boundary of the belt and the centre of symmetry offset are known signatures of planet-mass bodies sculpting a debris disk. Fomalhaut's centre of symmetry offset may originate from an eccentric companion that forces the orbital elements of minor bodies

into pericentre alignment^{21,22}. The net result is a torus of planetesimals with one side uniformly closer to the star. A ~ 2 AU centre of symmetry offset is also inferred for the dust belt surrounding the star HR 4796A, based on an observed excess of warm grains on one side of the star²¹. A key question is if the sharp inner edge to Fomalhaut's belt is due to planetary resonances, or to ejection. Models of our Kuiper belt dust distribution indicate that grains accumulate in mean motion resonances with Neptune, producing ring-like features²³. These models also show that at ~ 10 AU radius, particles of all sizes are ejected by dynamical interactions with Jupiter and Saturn. We expect that future observations will constrain these possibilities by directly detecting planets around Fomalhaut, or by finding azimuthal dust features that will further elucidate the dynamics of the system.

The emerging picture of Fomalhaut is that of a dust belt with prominent features attributable to a planetary system. Other possible perturbers such as a low-mass companion outside the belt may also modify the structure of the belt. The ~ 200 Myr age of Fomalhaut²⁴ is one order of magnitude older than other optically detected debris disk systems such as β Pic⁶ and AU Mic⁷, and an order of magnitude younger than our Kuiper belt (Table 1). Further understanding of Fomalhaut's belt may elucidate the epoch of evolution that for our Solar System corresponds to the radial migration of planets owing to the gravitational scattering of minor bodies, the recent formation of Pluto-sized objects at the outer edge of the system, the inward delivery of primordial icy material to the surfaces and atmospheres of terrestrial planets, and the formation of the Kuiper belt and the Oort cloud.

METHODS

Because Fomalhaut is one of the brightest stars in the sky (V-band magnitude $m_V = 1.16$ mag), we used the ACS coronagraphic camera with a $1.8''$ diameter occulting spot to artificially eclipse the star. A red broadband filter (F814W, central wavelength $\lambda_c = 833$ nm, width $\Delta\lambda = 251$ nm), comparable to the Johnson-Cousins I-band, was chosen for observations at three epochs in 2004 (17 May, 2 August, 27 October) and with different telescope roll angles at each epoch (the +Y-axis of the detector had PA 67.6° , 93.5° and 222.6° , respectively). The F814W data have a cumulative integration time of ~ 80 min. Additional follow-up data in the third epoch were obtained with the F606W filter ($\lambda_c = 591$ nm, $\Delta\lambda = 234$ nm), comparable to the Johnson-Cousins V-band, with +Y-axis PA 222.6° , 226.6° and 230.6° within the third epoch. For all epochs and wavelengths, the main reference star for subtraction of the point-spread function (PSF) was HD 172167, which was also observed with a wide range of position angles, but always in telescope orbits immediately before or after the Fomalhaut observations. In the third epoch, a second PSF reference star (HD 210418) was also observed with the F606W filter. Short integrations produced a set of images where the unsaturated PSF core is visible at high signal to noise behind the occulting spot, giving the location of the stellar centroid to a fraction of a pixel (< 25 mas). The stellar full-width at half-maximum is 63 mas (0.5 AU) in F606W and 72 mas (0.6 AU) in F814W.

Data reduction included the standard pipeline processing from the HST archive that produces bias-subtracted and flat-fielded image files. Bad pixels were replaced by interpolation using neighbour pixel values. Each Fomalhaut PSF was then subtracted iteratively using the reference star data. Registration was typically sensitive to 0.1 pixel relative shifts. The images were then corrected for distortion, resulting in images with 25 mas per pixel, registered to a common orientation, and median-combined. To test for nebulosity in our PSF reference star images, we performed the same data reduction steps, including registration to a common orientation, and further binned the data 16×16 pixels, with no detection of nebulosity in our PSF reference star fields. The Fomalhaut belt is detected at each epoch of observation, as are several background stars. To improve the signal-to-noise in the belt, we combined our F606W and F814W data using a weighted average $[(0.69 \times F606W + F814W)/2]$. For the combined data we use zero point equal to 24.4 mag. Assuming that Fomalhaut $m_V - m_I = 0.09$ mag, the stellar magnitude in the combined data is $m_{F606W+F814W} = 1.12$ mag. Photometry on the belt was conducted using a circular aperture with diameter 1.25 arcsec centred on the brightest points along the belt. The images shown in Figs 1 and 2 represent the combined data and they are not smoothed or binned.

The model belt shown in Figs 1b and 3 has a volume number density distribution that varies as a function of radius and height above the midplane¹³.

The vertical number density away from the midplane decreases exponentially, consistent with that derived for the edge-on disks β Pic and AU Mic^{25,26}. The scale height flares as a power-law function of radius with exponent equal to 0.5. As described in the text, the model parameters (inner and outer radius, inclination to the line of sight, position angle, phase function asymmetry parameter, radial number density index, scale height, and flare index) were adjusted to provide a satisfactory subtraction of the ring in Fig. 1, and an approximate fit to the radial cut shown in Fig. 3. As the belt may not have a uniform physical width around its circumference, the model belt chosen here is non-unique and may only apply to the section of the belt sampled in Fig. 2b. In Fig. 3, least-squares exponential and power-law fits to the belt gradient in the inner edge (120–140 AU) are proportional to $\exp(0.08r)$ and $r^{10.9}$, respectively, where r is in AU. The outer edge gradient (140–158 AU) is proportional to $\exp(-0.03r)$ and $r^{-4.6}$, respectively.

Received 19 November 2004; accepted 28 March 2005.

- Backman, D. E. & Paresce, F. in *Protostars and Protoplanets III* (eds Levy, E. H. & Lunine, J. I.) 1253–1304 (Univ. Arizona Press, Tucson, 1993).
- Roques, F., Scholl, H., Sicardy, B. & Smith, B. A. Is there a planet around beta Pictoris? Perturbations of a planet on a circumstellar dust disk. I: The numerical model. *Icarus* **108**, 37–58 (1994).
- Liou, J.-C. & Zook, H. A. Signatures of giant planets imprinted on the Edgeworth-Kuiper Belt dust disk. *Astron. J.* **118**, 580–590 (1999).
- Ozernoy, L. M., Gorkavyi, N. N., Mather, J. C. & Taidakova, T. A. Signatures of extrasolar planets on dust debris disks. *Astrophys. J.* **537**, L147–L151 (2000).
- Greaves, W. S. et al. Structure in the epsilon Eridani debris disk. *Astrophys. J.* **619**, L187–L190 (2005).
- Smith, B. A. & Terrile, R. J. A circumstellar disk around Beta Pictoris. *Science* **226**, 1421–1424 (1984).
- Kalas, P., Liu, M. C. & Matthews, B. C. Discovery of a large dust disk around the nearby star AU Microscopii. *Science* **303**, 1990–1993 (2004).
- Aumann, H. H. IRAS observations of matter around nearby stars. *Publ. Astron. Soc. Pacif.* **97**, 885–891 (1985).
- Gillet, F. in *Light on Dark Matter* (ed. Israel, F. P.) 61–69 (Reidel, Dordrecht, 1986).
- Holland, W. S. et al. Submillimeter images of dusty debris around nearby stars. *Nature* **392**, 788–791 (1998).
- Dent, W. R. F., Walker, H. J., Holland, W. S. & Greaves, J. S. Models of the dust structures around Vega-excess stars. *Mon. Not. R. Astron. Soc.* **314**, 702–712 (2000).
- Smart, W. M. On the derivation of the elements of a visual binary orbit by Kowalsky's method. *Mon. Not. R. Astron. Soc.* **90**, 534–538 (1930).
- Kalas, P. & Jewitt, D. The detectability of Beta Pic-like circumstellar disks around nearby main sequence stars. *Astron. J.* **111**, 1347–1355 (1996).
- Clampin, M. et al. Hubble Space Telescope ACS coronagraphic imaging of the circumstellar disk around HD 141569. *Astrophys. J.* **126**, 358–392 (2003).
- Fixsen, D. J. & Dwek, E. The zodiacal emission spectrum as determined by COBE and its implications. *Astrophys. J.* **578**, 1009–1014 (2002).
- Stapelfeldt, K. R. et al. First look at the Fomalhaut debris disk with the Spitzer Space Telescope. *Astrophys. J. Suppl.* **154**, 439–442 (2004).
- Artymowicz, P. & Clampin, M. Dust around main-sequence stars: Nature or nurture by the interstellar medium? *Astron. J.* **490**, 863–878 (1997).
- Wyatt, M. C. & Dent, W. R. F. Collisional processes in extrasolar planetesimal discs — dust clumps in Fomalhaut's debris disc. *Mon. Not. R. Astron. Soc.* **334**, 589–607 (2002).
- Holland, W. S. et al. Submillimeter observations of an asymmetric dust disk around Fomalhaut. *Astrophys. J.* **582**, 1141–1146 (2003).
- Chiang, E. I. & Goldreich, P. Apse alignment of narrow eccentric planetary rings. *Astrophys. J.* **540**, 1084–1090 (2000).
- Wyatt, M. C. et al. How observations of circumstellar disk asymmetries can reveal hidden planets: Pericenter glow and its application to the HR 4796 disk. *Astrophys. J.* **527**, 918–944 (1999).
- Kuchner, M. J. & Holman, M. J. The geometry of resonant signatures in debris disks with planets. *Astrophys. J.* **588**, 1110–1120 (2003).
- Moro-Martín, A. & Malhotra, R. A study of the dynamics of dust from the Kuiper Belt: Spatial distribution and spectral energy distribution. *Astron. J.* **124**, 2305–2321 (2002).
- Barrado y Navascués, D. The Castor moving group. The age of Fomalhaut and Vega. *Astron. Astrophys.* **339**, 831–839 (1998).
- Kalas, P. & Jewitt, D. Asymmetries in the Beta Pictoris dust disk. *Astron. J.* **110**, 1008–1017 (1995).
- Krist, J. et al. Hubble Space Telescope ACS coronagraphic imaging of the AU Microscopii debris disk. *Astron. J.* **129**, 1008–1017 (2005).
- Luu, J. X. & Jewitt, D. Kuiper Belt Objects: Relics from the accretion disk of the Sun. *Annu. Rev. Astron. Astrophys.* **40**, 63–101 (2002).
- Kelsall, T. et al. The COBE diffuse infrared background experiment search for the cosmic infrared background. II. Model of the interplanetary dust cloud. *Astrophys. J.* **508**, 44–73 (1998).

Acknowledgements This research is based on observations with the NASA/ESA Hubble Space Telescope, which is operated by the Association of Universities for Research in Astronomy. P.K. acknowledges support from the Space Telescope Science Institute and NASA's Origins of Solar Systems programme. P.K. and J.R.G. also thank the NSF Center for Adaptive Optics, managed by the University of California, Santa Cruz.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.K. (kalas@astron.berkeley.edu).

Timescales of shock processes in chondritic and martian meteorites

P. Beck¹, Ph. Gillet¹, A. El Goresy² & S. Mostefaoui²

The accretion of the terrestrial planets from asteroid collisions and the delivery to the Earth of martian and lunar meteorites has been modelled extensively^{1,2}. Meteorites that have experienced shock waves from such collisions can potentially be used to reveal the accretion process at different stages of evolution within the Solar System. Here we have determined the peak pressure experienced and the duration of impact in a chondrite and a martian meteorite, and have combined the data with impact scaling laws to infer the sizes of the impactors and the associated craters on the meteorite parent bodies. The duration of shock events is inferred from trace element distributions between coexisting high-pressure minerals in the shear melt veins of the meteorites. The shock duration and the associated sizes of the impactor are found to be much greater in the chondrite (~1 s and 5 km, respectively) than in the martian meteorite (~10 ms and 100 m). The latter result compares well with numerical modelling studies of cratering on Mars, and we suggest that martian meteorites with similar, recent ejection ages (10⁵ to 10⁷ years ago)³ may have originated from the same few square kilometres on Mars.

The peak pressure is usually established by comparing the deformation and transformation features observed in natural meteorites with those induced in experimentally shocked samples^{4,5}. Alternatively, peak pressures can be inferred using a combination of mineralogical phase transformations and high-pressure crystallization occurring within shock melt veins (SMV)^{6–8}. The SMVs result from high strain rates and frictional heating that are present during the pressure increase^{9–11}. The peak pressure developed within the SMVs is hydrostatic and it is constrained to be identical with their surroundings owing to stress continuity. The temperature within the SMVs is significantly higher (by up to 2,500 K) than in the whole rock, leading to the formation of high-pressure assemblages and chemical diffusion that do not occur in the bulk. Here we have measured the concentrations of trace elements (Mn, Ca, Ba, Cs, Sr) in high-pressure minerals that crystallized during the peak pressure pulse. The element concentrations were obtained on a nanometre scale using secondary ion mass spectrometry (nano-SIMS). The partitioning of these elements between coexisting phases was then used to constrain the duration of the peak pressure pulse. We tested this approach on meteorite samples (Tenham L6 chondrite; Zagami shergottite) for which the shock histories were already well-characterized^{9,10,12–14}.

According to the current shock classification scale⁵, the Tenham meteorite experienced a peak pressure between 55 and 90 GPa. However, it has been shown recently using similar criteria that the peak pressure is more likely to lie in the 25–45 GPa range¹³. In addition, the mineralogy within the SMVs in Tenham is dominated by the high-pressure polymorphs of (Mg,Fe)₂SiO₄ and (Mg,Fe)SiO₃ (refs 12, 14). Plagioclase transformed to hollandite during the shock⁷, and the assemblage magnesio-wüstite + majorite^{6,15} crystallized from

the melt. These observations indicate that the pressure and temperature conditions developed within the SMV were 23–25 GPa, consistent with the lower bound on the peak shock pressure estimated from textural evidence, and around 2,500 K (±100 K). The replacement of olivine by ringwoodite lamellae forming along crystallographic orientations (Fig. 1a) is one of the characteristic high-pressure transformation features in the SMVs. Quantitative trace element mapping shows that the Ca contents are nearly identical in both phases, but that ringwoodite is depleted in Mn with respect to olivine with an apparent partition coefficient (Fig. 1b) $D_{\text{Ring/Ol}}^{\text{Mn}} = 0.75 \pm 0.04$. Ca and Mn are present as trace elements within the two phases, and their exchange is controlled by diffusion, as shown by the concentration profile across ringwoodite lamellae (Fig. 1e). Mn has had time to diffuse during the peak shock pressure timescale, whereas Ca has remained immobile. Using known diffusion rates, it is possible to calculate diffusion lengths for the two elements, as a function of peak shock pulse duration and cooling rate (Fig. 2). Results are presented for lengths of 1, 1.2 and 1.5 μm, corresponding to the observed widths of ringwoodite lamellae.

The possible solution field for the peak shock pressure duration is delimited by the Ca 1-μm curve, because the diffusion rate of Ca is too small to have resulted in significant migration, and by the Mn 1.5-μm curve, because Mn diffusion was observed to have occurred over this distance (Table 1 and Fig. 1b). These curves can be compared with the width of ringwoodite lamellae computed from the experimentally determined growth rate of ringwoodite within olivine¹⁶, recently used to estimate the shock duration in another chondrite¹⁷. The growth curves are located within the solution field inferred from the diffusion curves of Mn and Ca. Regardless of the cooling rate, the peak shock pressure duration (Δt) must have lasted between 0.2 s and 5 s. A more precise estimate for Δt (0.2–1 s) is obtained by evaluating the cooling rate of the vein (Fig. 2).

In Zagami, a basaltic shergottite, the SMV mineralogy is dominated by K- and (Ca,Na)-hollandite, stishovite, omphacite, akimotoite and amorphous grains of former silicate perovskite⁹. In the SMV studied here, K-hollandite is associated with stishovite (Fig. 1c), interpreted as a liquidus assemblage⁹. The matrix is mostly composed of quenched glass associated with the liquidus majorite + magnesio-wüstite assemblage. From these observations, the peak shock pressure and temperature are estimated to be 22–23 GPa and 2,400–2,500 K (ref. 18). The shock duration was then obtained from studying trace-element concentrations determined for 10-μm-sized liquidus aggregates of K-hollandite surrounded by stishovite grains using nano-SIMS (Fig. 1d). The concentrations of Cs, Ba and Rb in K-hollandite are higher than those for the host rock, or those measured outside the SMVs in maskelynite grains (Table 2). Surrounding minerals are also largely depleted in these elements compared with K-hollandite (Fig. 1d).

The observations are consistent with trace-element fractionation

¹Laboratoire de Sciences de la Terre, CNRS UMR 5570, Ecole Normale Supérieure de Lyon et Université Lyon I, 46 allée d'Italie, 69364 Lyon Cedex 7, France. ²Max-Planck-Institut für Chemie, D-55128 Mainz, Germany.

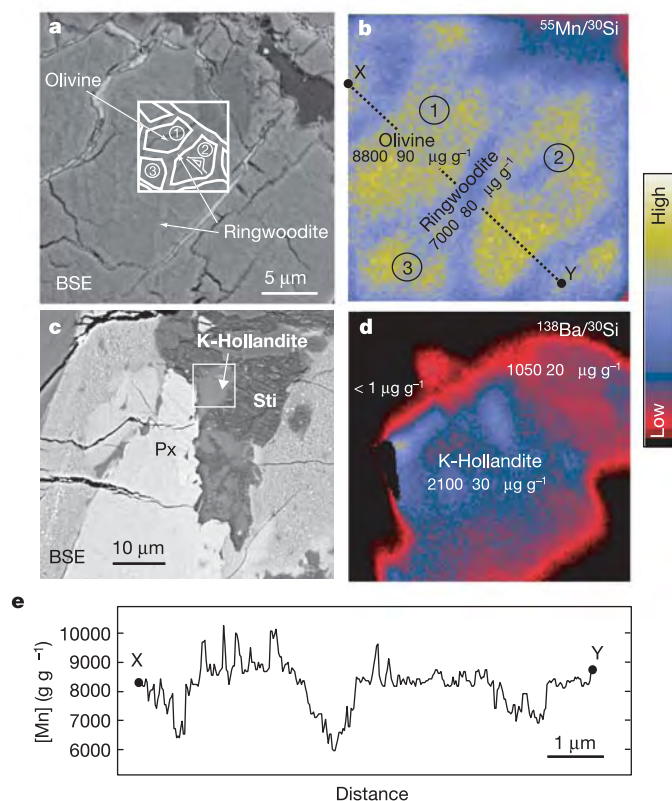


Figure 1 | Trace element maps of high-pressure minerals in Zagami and Tenham meteorites. **a**, Back-scattered electron (BSE) image of an area in a shock melt vein of the Tenham meteorite where ringwoodite (indicated brighter grey) laths have been formed in an olivine grain (darker grey). Both phases have been identified by Raman spectroscopy⁷. They have sizes ranging between 1 and 1.5 μm . The numbers refer to the olivine zones also reported on the chemical $^{55}\text{Mn}/^{30}\text{Si}$ map (**b**). The white square corresponds to the zone where the trace-element mapping was performed. **b**, Corresponding $^{55}\text{Mn}/^{30}\text{Si}$ elemental image measured with a Cameca NanoSIMS 50 (Max-Planck-Institut für Chemie, Mainz). Ringwoodite is clearly depleted in Mn with respect to olivine. The NanoSIMS instrument is characterized by high spatial resolution (down to 50 nm), high sensitivity, and the simultaneous detection of up to six isotopes. In this study, positive secondary ions of ^{23}Na , ^{24}Mg , ^{30}Si , ^{39}K , ^{44}Ca , ^{48}Ti , ^{52}Cr , ^{55}Mn , ^{27}Al , ^{85}Rb , ^{133}Cs and ^{138}Ba were mapped in combined and multi-detection modes at a mass resolution $m/\Delta m$ of $\sim 4,500$ (^{30}Si), sufficient to separate most isobaric interferences. The ^{44}Ca was used in order to correct for the contribution of ^{48}Ca to the ^{48}Ti signal. Using primary current conditions (~ 5 – 10 pA on the sample surface), the primary ion beam of O^+ was focused onto the sample in spots ~ 500 nm in size. Ion image sequences of 256×256 and 128×128 pixels were acquired by scanning areas of 5×5 to $15 \times 15 \mu\text{m}^2$ with counting times of ~ 50 to 150 ms pixel $^{-1}$. NanoSIMS images are quantitative, and permit the calculation of Mn concentrations in both phases (Table 1) by image processing. The mean Mn concentrations are indicated on the olivine and ringwoodite zones. **c**, BSE image showing the K-hollandite aggregate in the martian meteorite Zagami in which trace-element (Cs, Ba and Rb) abundances were measured (Table 2). The K-hollandite aggregate is surrounded by the high-pressure SiO_2 -polymorph stishovite. Both phases have been identified by Raman spectroscopy and electron microprobe. **d**, NanoSIMS $^{138}\text{Ba}/^{30}\text{Si}$ elemental image of the K-hollandite aggregate in Zagami. The Ba concentrations (Table 2) in the grain are high with respect to the surrounding material, in agreement with a pronounced Ba migration during the shock. Similar measurements have been carried out for two other trace elements, Cs and Rb. **e**, Mn concentration profile measured across coexisting ringwoodite and olivine. The Mn concentration drop between ringwoodite and olivine is not as sharp as the olivine/ringwoodite interface is. This observation is consistent with a diffusion of Mn between olivine and ringwoodite. The location of the measured X–Y profile is shown in **b**.

between silicate melt and K-hollandite aggregates during the high pressure shock event. The crystallization of the K-hollandite can begin as soon as the shock-induced silicate melt appears and the pressure attains 23 GPa. At the end of the peak shock pulse, as the

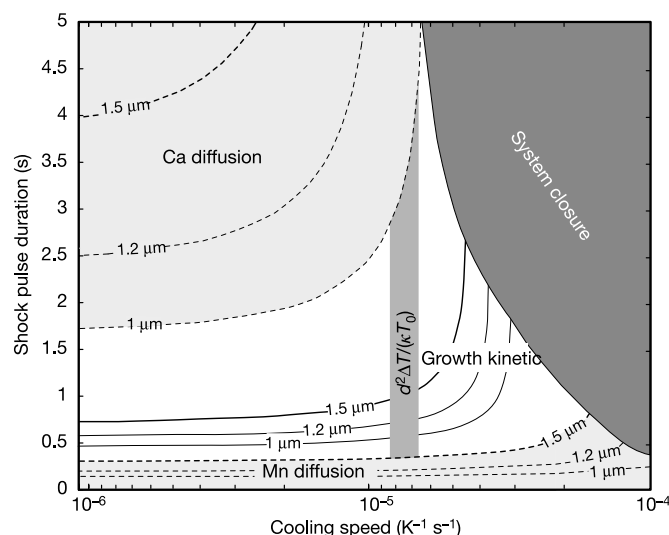


Figure 2 | Peak shock pressure duration as a function of the cooling speed for the Tenham meteorite. Three sets of curves are drawn, corresponding to the three methods used to constrain the shock pulse duration: Mn diffusion, Ca diffusion, and growth kinetics of ringwoodite laths within olivine¹⁰. The temperature dependence of both phenomena implies the knowledge of the cooling history of the system. Cooling was chosen to follow $1/T = at + (1/T_0)$ (ref. 26), where $a = d(1/T)/dt$ is the cooling speed and $T_0 = 2,500$ K is the starting temperature. Isoleths for Mn and Ca diffusion are calculated using:

$$x(t) = \sqrt{D_0} \exp \left[-\frac{aE_a}{2R} (t - t_0) \right]$$

where D_0 is the diffusion coefficient at T_0 , R the gas constant, E_a the activation energy, and t_0 the initial time. Both relations enable the calculation of the expected lamella width as a function of the shock pulse duration and the cooling speed. Curves corresponding to $x = 1$, 1.2 and 1.5 μm are shown for both Mn and Ca. The domain of solution for the shock pulse duration is delimited by the Mn 1.5- μm and Ca 1- μm curves, because on the one hand, Mn has diffused over this distance to form the 1.5- μm -sized laths, and on the other hand, Ca has not diffused over distances greater than 1 μm , though estimation from an elastic model²⁷ gives $D_{\text{Ring/Ol}}^{\text{Ca}} = 0.6$. A further refinement of this solution domain is provided by the observed size of the ringwoodite laths computed from the interface-controlled growth equation¹⁷ $v = k_0 \text{Exp}(-\Delta G_a/RT)(1 - \exp(\Delta G_r/RT))$, where v is the growth rate, R the gas constant, T the temperature in K, k_0 a constant, and ΔG_a the activation energy of the transformation. Values of k_0 and ΔG_a were taken from kinetic experiments on the intracrystalline growth of ringwoodite within olivine¹⁶. The free energy change ΔG_r of the olivine to ringwoodite transformation was calculated for each time step using the thermodynamic data of ref. 28. The chosen thermal history is the same as the one used for diffusion calculation, with a peak temperature of $T_0 = 2,500$ K. The pressure is taken as $P = 24$ GPa. Additional calculations using a linear and exponential cooling rate give very similar results. The right zone of the diagram (system closure) is the field for which the duration of the shock pulse is too short to reproduce the observed sizes of the ringwoodite laths, or for which the cooling is too fast to permit enough elemental transport of Ca and Mn. The typical characteristic cooling speed in the Tenham shock melt vein (vertical shaded line) can be estimated by a dimension analysis leading to:

$$a = \frac{d(1/T)}{dt} = -\frac{1}{T^2} \frac{dT}{dt} = -\frac{1}{T_0^2} \frac{\Delta T}{\Delta t} = -\frac{1}{T_0^2} \frac{\kappa \Delta T}{d^2}$$

where κ is the thermal diffusivity (typical value $10^{-6} \text{ m}^2 \text{ s}^{-1}$), d the half-width of the studied shock vein ($\sim 300 \mu\text{m}$) and ΔT the temperature variation (about $-2,000$ K).

pressure drops below 23 GPa, the vein remains molten and lower pressure mineral phases are formed⁹. The time required for trace elements to diffuse into the K-hollandite thus provides a means of calculating a maximum value for the duration (Δt) of the peak shock pressure. The amount of each trace element within K-hollandite aggregates is calculated from the measured concentrations, assuming a particle spherical in shape with the observed grain radius $r = 4.1 \mu\text{m}$. The starting chemical composition of the melt is assumed to be that of the whole rock, as the melting process is shear-induced and the SMVs cut across the whole sample. The corresponding spherical volumes (radii R_i) of melt fractions containing the same amount of Cs, Ba and Rb as the K-hollandite aggregate are $R_i = 14.7, 16.3$ and $17.6 \mu\text{m}$, respectively. The diffusion length for each element to enter the K-hollandite aggregates is then $(R_i - r)$. The time required for these elements to diffuse over the distance $(R_i - r)$ is calculated from $t = (R_i - r)^2/D_{i(T)}$, where $D_{i(T)}$ is the temperature-dependent diffusion coefficient of the element in the melt (Table 2). The crystallization temperature of K-hollandite from the high-pressure melt is constrained by the phase diagram of basaltic rocks¹⁸ ($T = 2,450 \pm 50 \text{ K}$, 23 GPa). Using these assumptions, the peak shock pressure duration calculated for all three trace elements

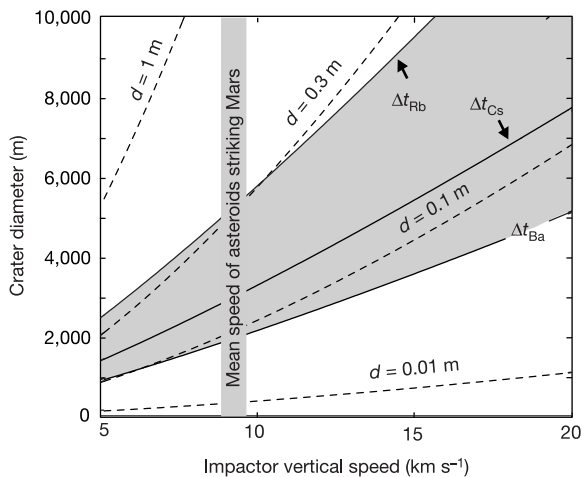


Figure 3 | Diameters of the impact craters on Mars as a function of the impactor vertical speed. The diameter is calculated using the pi-scaling relation¹⁹:

$$D_{\text{crater}} = 1.8 \times \rho_{\text{proj}}^{0.11} \times \rho_{\text{target}}^{-1/3} \times g^{-0.22} \times D_{\text{proj}}^{0.13} \times W^{0.22} \quad (1)$$

where $D_{\text{proj}} = \tau \times v_{\text{proj}}$. W is taken as pure kinetic energy, $\rho_{\text{proj}} = 3,700 \text{ kg m}^{-3}$ is the density of the impactor, $\rho_{\text{target}} = 2,800 \text{ kg m}^{-3}$ is the density of the basaltic target (corresponding to the martian basaltic crust, and the Zagami meteorite), and $g = 3.72 \text{ m s}^{-2}$, the acceleration due to gravity on Mars. We note that this relation is independent of the peak shock pressure. Dashed lines represent calculated iso-size curves of the ejected spalls (martian meteorite)^{19,29}. The mean size of the spalls is given by:

$$l = \frac{\gamma \times D_{\text{proj}}}{\rho_{\text{target}} v_{\text{impactor}}^{2/3} v_{\text{proj}}^{4/3}}$$

where $\gamma = 170 \text{ MPa}$ is the dynamic tensile stress of fresh gabbros or basalts³⁰, and v_{proj} the speed of the ejected spall, taken as the martian escape velocity. For the special case of Zagami and related shergottites of similar ejection ages, the expected curves relating the crater size to the impactor velocity is obtained by introducing in equation (1) the shock duration τ inferred from the Rb, Cs and Ba diffusion data. This relation is bracketed by the lowest duration time calculated from the Ba data and the highest shock duration determined from the Rb data (grey shaded area between curves). As the mean speed of present Mars-crossing asteroids is 9.3 km s^{-1} (vertical grey bar, ref. 20), the Zagami meteorite should come from a source crater of diameter 1,500–4,000 m. The expected size of the ejected spalls lies between a few centimetres and a few tens of centimetres, in excellent agreement with the actual size distribution of shergottites.

is of the order of $\sim 10 \text{ ms}$; $(\Delta t)_{\text{Ba}} = 10.0 \text{ ms}$, $(\Delta t)_{\text{Cs}} = 15.4 \text{ ms}$ and $(\Delta t)_{\text{Sr}} = 30.8 \text{ ms}$.

The shock pulse durations in Zagami and Tenham are $\sim 0.01 \text{ s}$ and $\sim 1 \text{ s}$ respectively, and these are correlated with the size of the impactors by the impact velocity¹⁹ (Fig. 3). For the martian meteorite with a typical impactor velocity of 10 km s^{-1} , the corresponding impactor size is 0.1 km , and for Tenham (velocity 5 km s^{-1}) it is approximately 5 km . Generally, the pi-scaling relation¹⁹ relates the size of the impactors to that of the craters formed on the target planet or body. Introducing in this relation an equation that links the shock duration τ , the impactor velocity v_{impactor} and the impactor size D_{impactor} ($D_{\text{impactor}} = \tau \times v_{\text{impactor}}$), we can compute the expected crater size as a function of the impactor speed (Fig. 3) for different shock durations. Using the shock durations in Zagami inferred from the Ba, Cs and Rb diffusion data (Fig. 3), and usual asteroid speeds ($5\text{--}20 \text{ km s}^{-1}$), the crater size should not exceed 10 km . Using the mean speed of Mars-crossing asteroids²⁰ (9.3 km s^{-1}) in these calculations, the crater size on Mars should range between $1,500 \text{ m}$ and $5,000 \text{ m}$ (corresponding to impactor sizes of 90 and 280 m). Recent calculations²¹ show that shergottites are probably ejected from small craters ($\sim 3 \text{ km}$). This result is consistent both with our shock duration data, and with the 'young' ejection ages determined for the shergottites (of the order of $10^5\text{--}10^7 \text{ years}$)³. The sizes of the ejected spalls¹⁹ (Fig. 3), that is, the sizes of martian meteorites arriving on Earth, range from a few centimetres to a few tens of centimetres, in excellent agreement with the actual size distribution of shergottites. Many shergottites possess similar ejection ages to Zagami, suggesting that they originate from the same shock event on Mars. Given the small crater size that we infer, it is further suggested that all these rocks of distinct petrology, magmatic ages ($180\text{--}350 \text{ Myr}$), and chemical and isotopic composition could have coexisted on Mars within a few square kilometres of each other, to a depth of a few hundred metres.

Tenham is representative of the most heavily shocked chondrites known. The age of its shock metamorphism is unknown. However, if it occurred early in the history of the Solar System, then the observed

Table 1 | Data from the Tenham L6 chondrite

		Ca	Mn
Map 1	Olivine	$60 \pm 15^*$	$0.880 \pm 0.009^\dagger$
	Ringwoodite	$78 \pm 17^*$	$0.7 \pm 0.008^\dagger$
	$D_{\text{Ring/Ol}}$	1.3 ± 0.6	0.79 ± 0.02
Map 2	Olivine	$490 \pm 33^*$	$0.906 \pm 0.007^\dagger$
	Ringwoodite	$443 \pm 39^*$	$0.656 \pm 0.006^\dagger$
	$D_{\text{Ring/Ol}}$	0.9 ± 0.1	0.75 ± 0.02
Log[D_0 ($\text{m}^2 \text{s}^{-1}$)] in olivine		-9.155 (ref. 22)	-7.167 (ref. 22)
E_a (kJ mol^{-1}) in olivine		176 (ref. 22)	218 (ref. 22)

Two elemental maps (Map 1 and Map 2) were recorded in two distinct olivine grains that were partly transformed into ringwoodite. For each map the mean concentration of Mn and Ca was calculated within adjacent olivine and ringwoodite. Map 1 corresponds to the Mn image presented in Fig. 1. The coefficients of diffusion for Mn and Ca in olivine used in the calculations are temperature-dependent, and follow the relation $D = D_0 \exp(-E_a/RT)$ with the values of D_0 and E_a shown in the table.

* Units, $\mu\text{g g}^{-1}$.

† Units, wt%.

Table 2 | Data from the Zagami shergottite

	Cs	Rb	Ba
Maskelynite	u.d.	$1.6 \pm 0.1^*$	$25.5 \pm 0.7^*$
Bulk rock	0.37^* (ref. 23)	5.32^* (ref. 23)	22.3^* (ref. 23)
K-Hollandite	$22 \pm 2^*$	$495 \pm 30^*$	$1,520 \pm 20^*$
Log[D_0 ($\text{m}^2 \text{s}^{-1}$)]	-2 (ref. 24)	-3.46 (ref. 24)	-2.097 (ref. 25)
E_a (kJ mol^{-1})	272 (ref. 24)	213 (ref. 24)	255 (ref. 25)

Cs, Rb and Ba concentrations measured in maskelynite and K-hollandite from the Zagami shergottite. Bulk rock concentrations are also presented. (u.d., under detection limit). As the Rb diffusion coefficient is not known for basaltic melt, it was chosen to be equal to the Sr diffusion coefficient.

* Units, $\mu\text{g g}^{-1}$.

difference in the duration of peak shock pressures between chondrites and martian meteorites would be consistent with the existence then of a greater number of large asteroids than in recent times.

Received 29 July 2004; accepted 24 March 2005.

1. Wetherill, G. W. Formation of the terrestrial planets. *Annu. Rev. Astron. Astrophys.* **18**, 77–113 (1980).
2. Gladman, B. J. & Burns, J. A. Mars meteorite transfer: Simulation. *Science* **274**, 161–162 (1996).
3. Nyquist, L. E. et al. Ages and geologic histories of Martian meteorites. *Space Sci. Rev.* **96**, 105–164 (2001).
4. Stöffler, D. et al. Shock metamorphism and petrography of the Shergotty achondrite. *Geochim. Cosmochim. Acta* **50**, 889–903 (1986).
5. Stöffler, D., Keil, K. & Scott, E. R. D. Shock metamorphism of ordinary chondrites. *Geochim. Cosmochim. Acta* **55**, 3845–3867 (1991).
6. Chen, M., Sharp, T. G., El Goresy, A., Wopenka, B. & Xie, X. The majorite-pyroxene-magnesiowüstite assemblage: constraints on the history of shock veins in chondrite. *Science* **271**, 1570–1573 (1996).
7. Gillet, P., Chen, C., Dubrovinsky, L. & El Goresy, A. Natural NaAlSi₃O₈-hollandite in the shocked Sixiangkou meteorite. *Science* **287**, 1633–1636 (2000).
8. Beck, P., Gillet, P., Gautron, L., Daniel, I. & El Goresy, A. A new natural high-pressure (Na,Ca)-hexaluminosilicate [(Ca_xNa_{1-x})Al_{3+x}Si_{3-x}O₁₁] in shocked Martian meteorites. *Earth Planet. Sci. Lett.* **219**, 1–12 (2004).
9. Langenhorst, F. & Poirier, J. P. "Eclogitic" minerals in a shocked basaltic meteorite. *Earth Planet. Sci. Lett.* **176**, 259–265 (2000).
10. Langenhorst, F. & Poirier, J. P. Anatomy of black veins in Zagami: clues to the formation of high-pressure phase. *Earth Planet. Sci. Lett.* **184**, 37–55 (2000).
11. Malavergne, V., Guyot, F., Benzerara, K. & Martinez, I. Description of new shock-induced phases in the Shergotty, Zagami, Nakhla and Chassigny meteorites. *Meteorit. Planet. Sci.* **36**, 1297–1305 (2001).
12. Putnis, A. & Price, G. D. High-pressure (Mg,Fe)₂SiO₄ phases in the Tenham chondritic meteorite. *Nature* **280**, 217–218 (1979).
13. Langenhorst, F., Joreau, P. & Doukhan, J. C. Thermal and shock metamorphism of the Tenham chondrite: A TEM examination. *Geochim. Cosmochim. Acta* **59**, 1835–1845 (1995).
14. Tomioka, N. & Fujino, K. Natural (Mg,Fe)SiO₃-ilmenite and perovskite in the Tenham meteorite. *Science* **277**, 1084–1086 (1997).
15. Agee, C. B., Li, J., Shannon, M. C. & Circone, S. Pressure-temperature phase-diagram for the Allende meteorite. *J. Geophys. Res. Solid Earth* **100**, 17725–17740 (1995).
16. Kerschhofer, L., Rubie, D. C., Sharp, T. G., McConnell, J. D. C. & Dupas-Bruzek, C. Kinetics of intracrystalline olivine-ringwoodite transformation. *Phys. Earth Planet. Inter.* **121**, 59–76 (2000).
17. Ohtani, E. et al. Formation of high-pressure minerals in shocked L6 chondrite Yamato 791384: constraints on shock conditions and parent body size. *Earth Planet. Sci. Lett.* **227**, 505–515 (2004).
18. Wang, W. & Takahashi, E. Subsolidus and melting experiment of a K-rich basaltic composition to 27 GPa: Implication for the behaviour of potassium in the mantle. *Am. Mineral.* **84**, 357–361 (1999).
19. Melosh, H. J. *Impact Cratering: A Geologic Process* (Oxford Univ. Press, Oxford, 1989).
20. Steel, D. Distributions and moments of asteroid and comet impact speeds upon the Earth and Mars. *Planet. Space Sci.* **46**, 473–478 (1998).
21. Head, J. N., Melosh, H. J. & Ivanov, B. A. Martian meteorite launch: High-speed ejecta from small craters. *Science* **298**, 1752–1755 (2002).
22. Jurewicz, A. J. G. & Watson, E. B. Cations in olivine. 2. Diffusion in olivine xenocrysts, with applications to petrology and mineral physics. *Contrib. Mineral. Petrol.* **99**, 186–201 (1988).
23. Barrat, J. A., Blichert-Toft, J., Nesbitt, R. W. & Keller, F. Bulk chemistry of Saharan shergottite Dar al Gani 476. *Meteorit. Planet. Sci.* **36**, 23–29 (2001).
24. Lowry, R. K., Henderson, P. & Nolan, J. Tracer diffusion of some alkali, alkaline-earth and transition element ions in a basaltic and an andesitic melt, and the implications concerning melt structure. *Contrib. Mineral. Petrol.* **80**, 254–261 (1982).
25. Freer, R. Diffusion in silicate minerals and glasses—a data digest and guide to the literature. *Contrib. Mineral. Petrol.* **76**, 440–454 (1981).
26. Dodson, M. H. Closure temperature in cooling geochronological and petrological systems. *Contrib. Mineral. Petrol.* **40**, 259–274 (1973).
27. Blundy, J. & Wood, B. Prediction of crystal-melt partition coefficients from elastic moduli. *Nature* **372**, 452–454 (1994).
28. Akaogi, M., Ito, E. & Navrotsky, A. Olivine-modified spinel-spinel transitions in the system Mg₂SiO₄-Fe₂SiO₄—Calorimetric measurements, thermochemical calculation, and geophysical application. *J. Geophys. Res. Solid Earth Planets* **94**, 15671–15685 (1989).
29. Warren, P. H. Lunar and Martian meteorite delivery services. *Icarus* **111**, 338–363 (1994).
30. Ahrens, T. J., Xia, K. & Coker, D. Shock-Compression of Condensed Matter. (eds Furnish, M. D., Thadhani, N. N. & Horie, Y.) 1393–1396 (American Institute of Physics, New York, 2001).

Acknowledgements We thank F. Albarède, B. Reynard and P. McMillan for reading and improving the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.B. (pbeck@ens-lyon.fr).

Structural signature of jamming in granular media

Eric I. Corwin¹, Heinrich M. Jaeger¹ & Sidney R. Nagel¹

Glasses are rigid, but flow when the temperature is increased. Similarly, granular materials are rigid, but become unjammed and flow if sufficient shear stress is applied. The rigid and flowing phases are strikingly different, yet measurements reveal that the structures of glass and liquid are virtually indistinguishable^{1,2}. It is therefore natural to ask whether there is a structural signature of the jammed granular state that distinguishes it from its flowing counterpart. Here we find evidence for such a signature, by measuring the contact-force distribution between particles during shearing. Because the forces are sensitive to minute variations in particle position, the distribution of forces can serve as a microscope with which to observe correlations in the positions of nearest neighbours. We find a qualitative change in the force distribution at the onset of jamming. If, as has been proposed^{3–9}, the jamming and glass transitions are related, our observation of a structural signature associated with jamming hints at the existence of a similar structural difference at the glass transition—presumably too subtle for conventional scattering techniques to uncover. Our measurements also provide a determination of a granular temperature that is the counterpart in granular systems to the glass-transition temperature in liquids.

Experiments on granular material provide a unique opportunity to measure quantities that are experimentally inaccessible in more microscopic systems. One such quantity is $P(F)$, the probability distribution of interparticle normal-force magnitudes¹⁰. For static, jammed granular systems, the shape of $P(F)$ has been studied by experiments and simulations^{8,10–21}. For frictional systems, $P(F)$ decays exponentially above the average force, $\langle F \rangle$, and has a plateau or small peak at force magnitudes below $\langle F \rangle$. This characteristic shape of $P(F)$ has become a signature of a jammed granular system¹⁹.

When the applied shear stress is raised above yield, granular material will flow²². Above yield, some simulations¹⁹ find the disappearance at high strain rates of the low-force peak in $P(F)$, and others²¹ suggest that even below yield $P(F)$ loses its peak parallel to the strain direction. However, another set of simulations finds a negligible change in $P(F)$ on crossing the jamming threshold²³. The measured distribution of impulses has shown²⁴ an increase in the number of low impulses at higher flow rates. Hard-sphere simulations²⁵ interpret this as due to an increase in the probability of high forces. These discrepancies demonstrate the current lack of consensus about any characteristic change in $P(F)$ as the system unjams.

Measurements of $P(F)$ in sheared packings require rapid-response transducers that measure forces on single beads *in situ*. We achieve these requirements by measuring $P(F)$ with a photoelastic plate at the bottom surface of a three-dimensional, cylindrical pack. This plate rotates the polarization of light in proportion to the applied local particle pressure. The position and magnitude of the local pressure is detected by a video camera that views the transducer through an analyser oriented to block any unrotated light (Fig. 1). A roughened piston applying a fixed normal load to the top surface is rotated at a constant rate so that the particle pressures on the bottom surface vary

with time. Extensive calibrations were performed to convert brightness and area values at each contact point into force magnitudes. This method is sensitive to the normal component of force, F , on each bead at the bottom surface of the pack^{13,16}. Earlier studies^{26,27} showed that measurements of the normal-force distribution at the surface of a pack match those measured within the bulk.

We can also measure the velocity profile of the flow along the bottom plate by tracking the position (without reference to brightness) of the spot that individual particles make on the plate as a function of time. The fixed external wall introduces strong radial shear at the container boundary, and the top surface applies a vertical shear-strain rate that increases linearly with distance, s , from the central axis. The insets to Fig. 2 show the in-plane radial shear-strain rate $\dot{\gamma} = s d\omega(s)/ds$ at the bottom surface as a function of radius, s , where $\omega(s)$ is the average angular velocity. For all heights and rotation rates, the radial strain rate is only significant near the outer edge and approaches zero near the central axis where the pack rotates rigidly. It is well fitted by an exponential decay $\dot{\gamma} \propto \exp[-(R_0 - s)/\lambda]$ with characteristic decay length $\lambda \approx 3\text{--}4$ bead diameters, where R_0 is the radius of the container.

Data were collected at the bottom surface in a region encompassing approximately 300 beads (bead diameter $d = 3.06 \pm 0.04$ mm), which stretched from the central axis to the outer edge of the pack at the container wall. A typical, hour-long data run provided $\sim 10^4$ frames for analysis, resulting in 3×10^6 distinct force measurements. This allowed determination of the time-averaged distribution, $P(f)$, over small regions of the bottom surface, where the variable f is the force magnitude, F , normalized by its average value across the plate, $\langle F \rangle$. Because the system is cylindrically symmetric, the shearing conditions are uniform in concentric rings around the axis. Thus, a normalized probability distribution of forces, $P_s(f)$, was calculated for each annulus as a function of radius, s (Fig. 2). In this way, we measure $P_s(f)$ within the shear band near the cell boundary as well as in the static region near the axis.

Figures 2a and b show $P_s(f)$ as a function of radius s for 10- and 20-particle-tall packs, respectively. We see a change in shape between the rigid regions near the axis and the sheared regions near the wall. We now show that there is an abrupt transition in $P(f)$ as the shear stress is increased above yield and that $P_s(f)$ from all sheared regions can be collapsed onto a single master curve. Figure 3a demonstrates this collapse for data from all heights, shear rates and applied loads measured.

Our data near the central axis match those of previous measurements taken for static systems of the same height^{13,16}, which at large forces show a slower than exponential distribution for short packs ($h < 20$) and a simple exponential distribution for taller packs ($h \geq 20$). We have also measured $P_s(f)$ under applied shear stress below yield and find that both the low-force behaviour (Fig. 3b) and the high-force behaviour (Fig. 3c) show no detectable change in the shape of the distribution as long as yield is not exceeded. However, at the outer edge of the pack, within the shear band, $P_{s>12d}(f)$ exhibits markedly different behaviour. First, it shows an enhancement for

¹James Franck Institute, Department of Physics, The University of Chicago, Chicago, Illinois 60637, USA.

forces smaller than the mean as compared to the static distribution (Fig. 3b). Second, it decays much more rapidly than an exponential.

The change in shape of the high-force part of $P_s(f)$ between the static and flowing regions can be quantified by measuring the curvature for a distribution of the general type $P_s(f) \approx \exp[-f^n]$. To extract the power n , we plot $P_s(f)$ on a (loglog)–(log) graph in the range $P_s(f) < 1$. In this plot, distributions representing the jammed phase would exhibit slope magnitudes no larger than unity ($n = 1$ for strictly exponential shapes and $n < 1$ for the shapes seen at small h). This is confirmed in Fig. 3c which includes data for all heights and shear stresses less than yield. In contrast, as shown in Fig. 3d, our data from the unjammed regime at all heights have slopes n significantly greater than unity.

The shear obtained from the velocity profiles (insets to Fig. 2) is in the radial direction. However, from the geometry of our system, we know that shear in the axial direction must exist as well. Since observed changes in $P_s(f)$ occur only near the container edge and not near its axis, we conclude that $P_s(f)$ is sensitive primarily to a shear band at the measurement surface oriented normal to the direction of force measurement. Little or no change is observed due to the vertical, or axial, shear above the measurement surface oriented parallel to the direction of force measurement.

To understand $P(f)$ in the flowing regime we start with the analytic

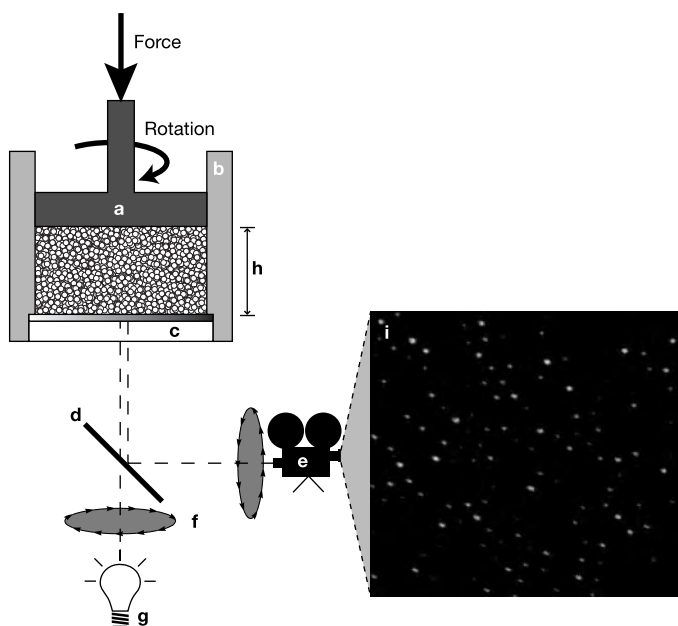


Figure 1 | Experimental set-up. Diagram of shear cell and force measurement apparatus. The bead pack has radius $R_0 \approx 21.5$ beads and heights (h) ranging from 6 to 20 beads, and is composed of soda-lime glass beads 3.06 ± 0.04 mm in diameter. Normal and shear stresses are applied to the bead pack through a roughened piston (a) with beads glued to it. For all measurements, a load of 1,380 N was applied by compressing a stiff spring to achieve a nearly constant loading condition. Shear force is applied by a 1/2-h.p. motor (torque 1,112 lb in) driving a 7:1 reduction sprocket to achieve rotation rates from 0.1 to 15 r.p.m. at the top of the bead pack. The bead pack is held in place by a confining cylinder (b) with smooth walls. The force transducer (c) sits underneath the bead pack, and consists of a 0.25-mm sheet of photoelastic polymer (PS-1E Vishay Measurements Group, Inc., Raleigh, NC) silvered on the top surface and bonded on the bottom to a 3/8-inch-thick clear glass plate. Beads pressing onto the top of the sheet create a local strain field, which rotates incident circularly polarized light. Illumination is provided by a 500-W slide projector (g) through a circular polarizer (f) and a half-silvered mirror (d). The force transducer is imaged by a digital video camera (Sony DCR VX-2000) (e) looking through an oppositely polarized circular polarizer (f). A sample image of forces taken with this apparatus is shown (i).

calculation of ref. 19, which predicts what $P(f)$ should be in an equilibrium system of interacting particles. Assuming that interparticle normal forces depend only on particle separation, we have $P(f)df = G(r)dr$ where $G(r)$ is the probability of finding particles separated by a distance between r and $r + dr$. In three dimensions, at asymptotically small r , $G(r) \propto r^2 \exp[-\beta V(r)]$, where $\beta = 1/(k_B T)$, T is the temperature, k_B is Boltzmann's constant, and $V(r)$ is the interparticle potential. Elastic spheres interact with the

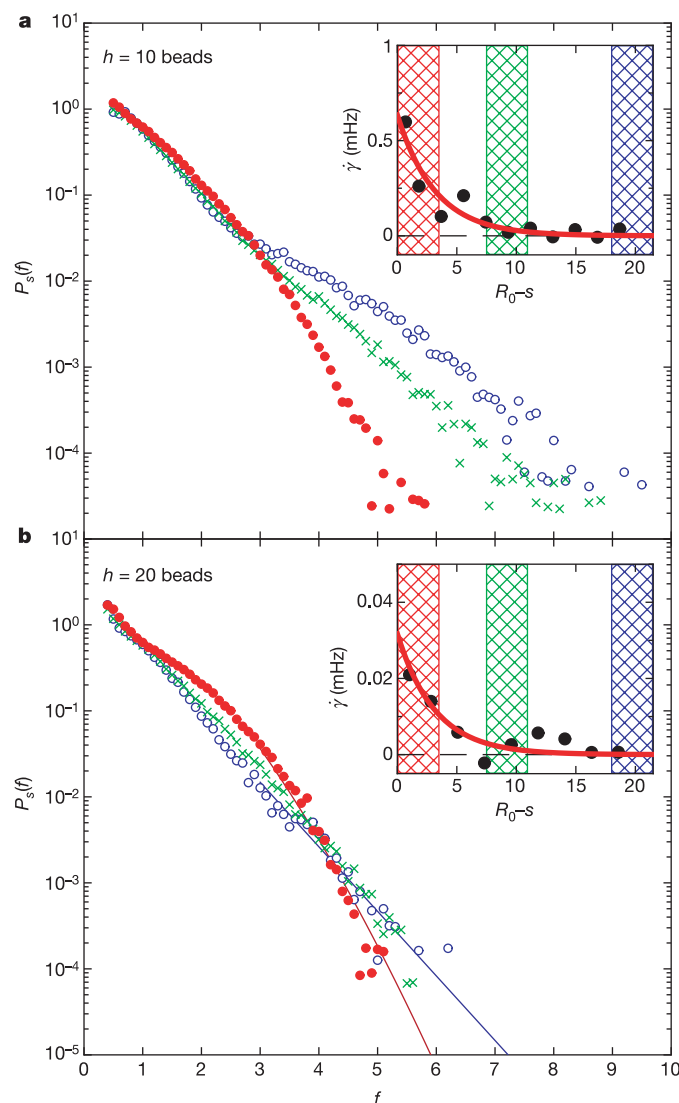


Figure 2 | Change of force probability distributions with local shear strain rate. Main panels, azimuthally averaged distributions, $P_s(f)$, evaluated inside annuli a distance, s , from the rotation centre are plotted for systems of two different filling heights, h . In both panels, data from three annuli are shown: $0 < s < 3.5$ (blue open circles), $10.5 < s < 14$ (green crosses), and $18.5 < s < 21$ (red filled circles). In the non-sheared regions near the centre, $P(f)$ exhibits the same shape as for a static pack (slower than exponential for $h = 10$ and nearly exponential for $h = 20$). In the shearing regions, $P(f)$ acquires the shape of an equilibrium distribution given by equation (1). Guides to the eye in the $h = 20$ graph are drawn to distinguish the data from the inner and outer annuli. Insets, in-plane shear strain rate $\dot{\gamma} = s d\omega(s)/ds$ as a function of distance ($R_0 - s$) from the container wall. The cross-hatched areas indicate the corresponding annular regions in the main panels. Packs rotating as rigid bodies correspond to $\dot{\gamma} = 0$ (dashed line). The decay of the shear strain rate away from the outer wall and towards the centre is well fitted by an exponential with characteristic length $\lambda = 3.2$ bead diameters for both $h = 10$ and $h = 20$ (solid lines). This behaviour is seen at all heights and all shear strain rates studied.

hertzian potential $V(r) \propto \Delta^{5/2}$, where $\Delta \equiv d - r$ and d is the particle diameter. In equilibrium we predict that for large forces

$$P(f) = \alpha \left[1 + f^{2/3} \frac{\langle \Delta \rangle^2}{d} \right]^2 \frac{1}{f^{1/3}} \exp \left[-\frac{\beta}{\beta_0} f^{5/3} \right] \quad (1)$$

where $\langle \Delta \rangle$ is the average deformation of a bead, $1/\beta_0$ is a temperature scale set by the average force per bead and the bead elastic modulus, and α is a normalization constant. Typical values of $\langle \Delta \rangle/d$ are about

3×10^{-5} so the first term in brackets can, to high accuracy, be replaced by unity. For very small forces, we can expand $G(r)$ around $r = d$. As long as there is any pressure in the system, G is a constant to leading order so we again obtain the low-force behaviour: $P(f) \propto f^{-1/3}$.

This model predicts that all data in the shear band at both high and low forces will follow a master curve if plotted as $P_s(f) T_{\text{eff}}^{1/5} / \alpha$ versus $f T_{\text{eff}}^{3/5}$, where $T_{\text{eff}} \equiv \beta_0 / \beta$ is a dimensionless effective temperature.

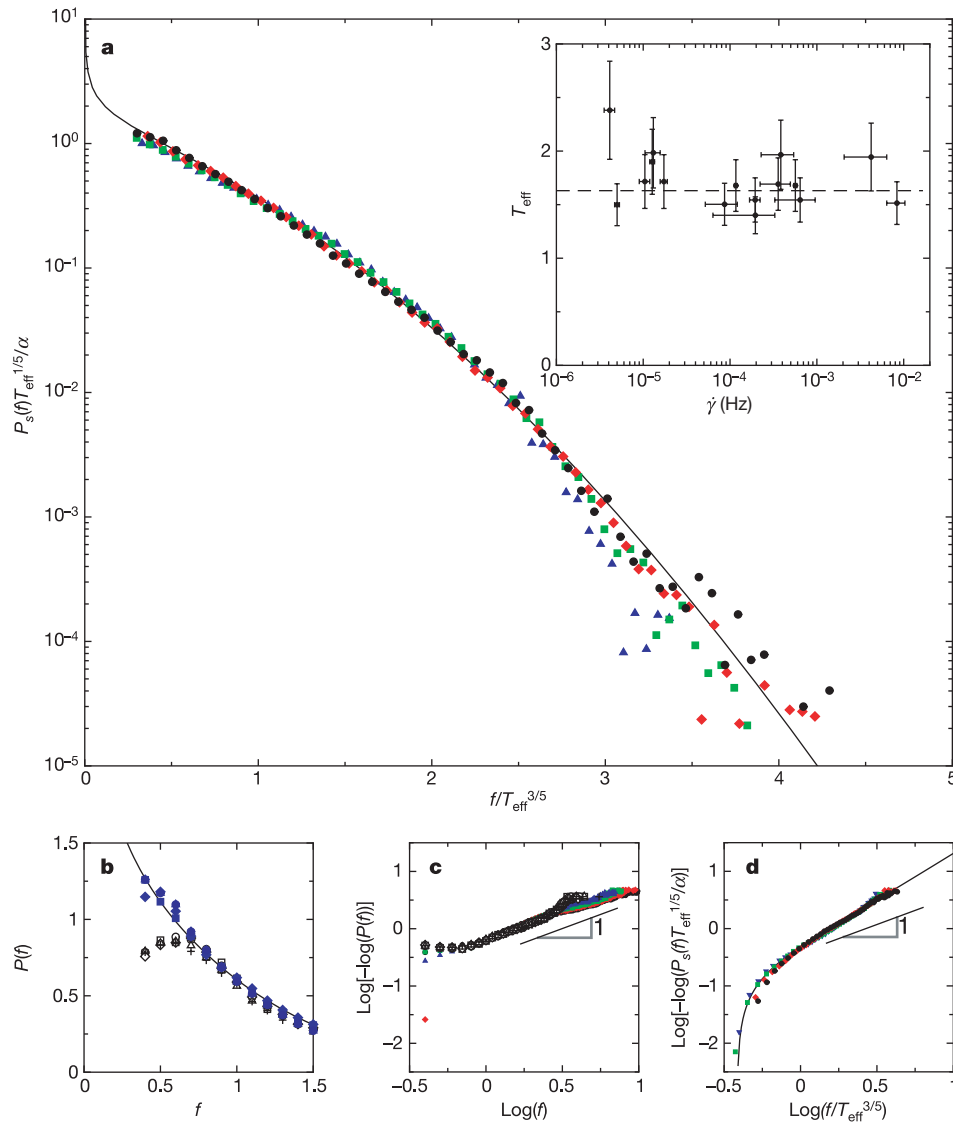


Figure 3 | Comparison between force probability distribution and equilibrium hertzian model. **a**, Scaled distributions $P_s(f) T_{\text{eff}}^{1/5} / \alpha$ in the sheared region as a function of scaled force magnitude $f T_{\text{eff}}^{3/5}$. Data from shearing regions (outer annulus, $18.5 < s < 21$) collapse onto a single master curve for all heights: $h = 6$ (black circles), 10 (red diamonds), 15 (green squares) and 20 (blue triangles). The solid line is the prediction of the hertzian model, equation (1). Inset, effective temperature, T_{eff} , obtained from the data collapse for packs of all heights, as a function of shear rate, $\dot{\gamma}$. The error bars in T_{eff} are associated with fitting our model to $P_s(f)$. The error bars in $\dot{\gamma}$ are the standard deviation of particle shear strain rates measured across the shearing annulus. Within experimental accuracy, there is no evolution of effective temperature with shear-strain rate over 3.5 orders of magnitude, from $\dot{\gamma} = 5 \times 10^{-6}$ Hz to 10^{-2} Hz. The dotted line shows the average value $\langle T_{\text{eff}} \rangle = 1.63$. The difference between the static and the sheared behaviour as well as the close agreement between sheared behaviour and the hertzian model is shown in greater detail for low forces in the linear-linear plot (**b**) and for asymptotically high forces in the log-log plots (**c** and **d**).

b, Enhancement of $P_s(f)$ at low forces in the shearing region as compared to a static material. Solid blue symbols are data from **a** (circles $h = 6$, diamonds $h = 10$, squares $h = 15$, triangles $h = 20$). This is contrasted with measurements from static packs (all with $h = 20$) at zero shear stress (open black squares), 1/4 yield stress (open black circles), 1/2 yield stress (open black diamonds), 3/4 yield stress (open black triangles), and just below yield (black crosses). The solid line shows the fit given by the hertzian contact model. **c**, Plot of $\log[-\log(P_s(f))]$ versus $\log(f)$ in the high-force regime $P_s(f) < 1$ for the inner, static region of sheared packs ($h = 6$ (black circles), 10 (red diamonds), 15 (green squares) and 20 (blue triangles)) and static packs ($h = 20$) experiencing shear stresses between zero up to just below yield (same symbols as **b**). **d**, Plot of $\log[-\log(P_s(f) T_{\text{eff}}^{1/5} / \alpha)]$ versus $\log(f T_{\text{eff}}^{3/5})$ in the high-force regime $P_s(f) < 1$ for the outer, shearing region (same data as in **a**). The solid line shows the fit given by the full hertzian contact model, equation (1), which has an asymptotic slope of $+5/3$. A slope of $+1$ corresponding to an exponential $P_s(f)$ is drawn for comparison in **c** and **d**.

Figure 3a demonstrates that data from multiple heights do indeed collapse onto a single curve in excellent agreement with equation (1). In the shearing region the flowing material thus behaves as if it were in equilibrium at temperature T_{eff} . This T_{eff} reflects the ability to explore phase space under shear.

We find that T_{eff} in the shearing region of our pack is insensitive to the measured shear rate at the bottom surface over a range from $\dot{\gamma} = 5 \times 10^{-6}$ to 10^{-2} Hz (Fig. 3 inset). The effective temperature is insensitive to pack height from $h = 6$ to 20 beads as well as to the driving angular velocity of the top plate. This flat behaviour of T_{eff} as a function of $\dot{\gamma}$ is similar to that seen in simulations by Ono *et al.*²⁸ of sheared systems where, at asymptotically slow strain rates, the effective temperature was interpreted as the temperature where the amorphous glass becomes a liquid. Ono *et al.* normalized their temperatures by the amount of energy necessary to move one particle past its neighbour. If we normalize our T_{eff} in the same way we find values between 0.001 and 0.002, consistent with 0.0015 as obtained in the Ono *et al.* simulations. The stated values are obtained from $T_{\text{eff}}/(\beta_0 \mu \langle F \rangle d)$ where μ is the coefficient of friction between glass spheres¹⁶ and β_0 is the energy stored in a bead compressed by a force $\langle F \rangle$ (ref. 29). Even if we allow μ to vary over a full range of reasonable values our effective temperatures will remain of the same order of magnitude.

Our results demonstrate a clear signature of the jamming/unjamming transition which manifests itself in the shape of the distribution of normal forces $P(f)$. Jammed packings are characterized by a distribution that decays exponentially at large forces. As recent simulations have shown, this shape reflects the non-equilibrium character of the jammed state¹⁸. By contrast, $P(f)$ in the flowing regime is well described by a model that assumes that the system is in equilibrium. Analysis of the shape of $P(f)$ in this regime gives both the functional form of the interparticle potential (in our case the hertzian potential, as appropriate for glass spheres at contact) and an effective granular temperature.

In molecular systems, measurements of structural changes typically focus on the pair correlation function, $g(r)$. A measurement of $P(f)$ has several advantages over measuring the pair correlation function. First, even a small polydispersity broadens $g(r)$ but leaves $P(f)$ unaffected since forces depend on compression from the point of contact rather than absolute interparticle distance. Second, even a tiny change in the interparticle spacing produces an extremely large change in the force. This makes the measurement of forces particularly sensitive to interparticle spatial correlations. The result of this paper on the signature of flowing versus jammed systems observed in $P(f)$ hints that such a signature might also exist at the glass transition. Moreover, the result that a flowing granular material can be approximately understood in terms of an equilibrium theory adds credence to the idea of a jamming phase diagram³. Although the region at large shear strain rate is flowing and so not in true equilibrium, it can, nonetheless, be described in terms of equilibrium distributions; that is, the shear rate can be thought of as simply producing an effective temperature which allows the material to explore phase space.

Received 31 January; accepted 24 April 2005.

1. Frick, B. & Richter, D. The microscopic basis of the glass-transition in polymers from neutron-scattering studies. *Science* **267**, 1939–1945 (1995).
2. Leheny, R. L. *et al.* Structural studies of an organic liquid through the glass transition. *J. Chem. Phys.* **105**, 7783–7794 (1996).

3. Liu, A. J. & Nagel, S. R. Nonlinear dynamics—Jamming is not just cool any more. *Nature* **396**, 21–22 (1998).
4. Liu, A. J., Nagel, S. R. (eds) *Jamming and Rheology. Constrained Dynamics on Microscopic and Macroscopic Scales* (Taylor and Francis, London, 2001).
5. Silbert, L. E., Ertas, D., Grest, G. S., Halsey, T. C. & Levine, D. Analogies between granular jamming and the liquid-glass transition. *Phys. Rev. E* **65**, 051307 (2002).
6. Coniglio, A., Fierro, A., Herrmann, H. J. & Nicodemi, M. (eds) *Unifying Concepts in Granular Media and Glasses* (Elsevier, Amsterdam, 2004).
7. Trappe, V., Prasad, V., Cipelletti, L., Segre, P. N. & Weitz, D. A. Jamming phase diagram for attractive particles. *Nature* **411**, 772–775 (2001).
8. Lacey, N. & Glotzer, S. C. Dynamical heterogeneity and jamming in glass-forming liquids. Preprint at (<http://ArXiv.org/cond-mat/0406451>) (2004).
9. Howell, D., Behringer, R. P. & Veje, C. Stress fluctuations in a 2D granular Couette experiment: A continuous transition. *Phys. Rev. Lett.* **82**, 5241–5244 (1999).
10. Liu, C. H. *et al.* Force fluctuations in bead packs. *Science* **269**, 513–515 (1995).
11. Howell, D. & Behringer, R. P. in *Powders and Grains 97* (eds Behringer, R. P. & Jenkins, J. T.) 337–340 (Balkema, Rotterdam, 1997).
12. Baxter, G. W. in *Powders and Grains 97* (eds Behringer, R. P. & Jenkins, J. T.) 345–348 (Balkema, Rotterdam, 1997).
13. Blair, D. L., Mueggenburg, N. W., Marshall, A. H., Jaeger, H. M. & Nagel, S. R. Force distributions in three-dimensional granular assemblies: Effects of packing order and interparticle friction. *Phys. Rev. E* **63**, 041304 (2001).
14. Løvøll, G., Måløy, K. J. & Flekkøy, E. G. Force measurements on static granular materials. *Phys. Rev. E* **60**, 5872–5878 (1999).
15. Makse, H. A., Johnson, D. L. & Schwartz, L. M. Packing of compressible granular materials. *Phys. Rev. Lett.* **84**, 4160–4163 (2000).
16. Mueth, D. M., Jaeger, H. M. & Nagel, S. R. Force distribution in a granular medium. *Phys. Rev. E* **57**, 3164–3169 (1998).
17. Silbert, L. E., Grest, G. S. & Landry, J. W. Statistics of the contact network in frictional and frictionless granular packings. *Phys. Rev. E* **66**, 061303 (2002).
18. O'Hern, C. S., Silbert, L. E., Liu, A. J. & Nagel, S. R. Jamming at zero temperature and zero applied stress: The epitome of disorder. *Phys. Rev. E* **68**, 011306 (2003).
19. O'Hern, C. S., Langer, S. A., Liu, A. J. & Nagel, S. R. Force distributions near jamming and glass transitions. *Phys. Rev. Lett.* **86**, 111–114 (2001).
20. Edwards, S. F. & Grinev, D. V. Statistical mechanics of granular materials: stress propagation and distribution of contact forces. *Granular Matter* **4**, 147–153 (2003).
21. Snoeijer, J. H., Vlucht, T. J. H., van Hecke, M. & van Saarloos, W. Force network ensemble: A new approach to static granular matter. *Phys. Rev. Lett.* **92**, 054302 (2004).
22. Conway, S. L., Shinbrot, T. & Glasser, B. J. A Taylor vortex analogy in granular flows. *Nature* **431**, 433–437 (2004).
23. Landry, J. W., Grest, G. S. & Plimpton, S. J. Forces in granular hopper flow. *Bull. Am. Phys. Soc.* **48**, 153 (2003).
24. Longhi, E., Easwar, N. & Menon, N. Large force fluctuations in a flowing granular medium. *Phys. Rev. Lett.* **89**, 045501 (2002).
25. Ferguson, A., Fisher, B. & Chakraborty, B. Impulse distributions in dense granular flows: Signatures of large-scale spatial structures. *Europhys. Lett.* **66**, 277–283 (2004).
26. Radjai, F., Roux, S. & Moreau, J. J. Contact forces in a granular packing. *Chaos* **9**, 544–550 (1999).
27. Bruijic, J., Edwards, S. F., Hopkinson, I. & Makse, H. A. Measuring the distribution of interdroplet forces in a compressed emulsion system. *Physica A* **327**, 201–212 (2003).
28. Ono, I. K. *et al.* Effective temperatures of a driven system near jamming. *Phys. Rev. Lett.* **89**, 095703 (2002).
29. Landau, L. D. & Lifshitz, E. M. *Theory of Elasticity* Ch. 1, 3rd edn, Sect. 9 (Butterworth-Heinemann, Oxford, 1986).

Acknowledgements We thank A. Bushmaker, X. Cheng, M. Möbius and N. Mueggenburg for help with the experiment, and S. Coppersmith, B. Chakraborty, A. Ferguson, A. J. Liu, N. Menon, C. S. O'Hern and L. Silbert for discussions about effective temperatures and force distributions. This work was supported by NSF-MRSEC, NSF-CTS and DOE.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.I.C. (ecorwin@uchicago.edu).

Contact force measurements and stress-induced anisotropy in granular materials

T. S. Majmudar¹ & R. P. Behringer¹

Interparticle forces in granular media form an inhomogeneous distribution of filamentary force chains. Understanding such forces and their spatial correlations, specifically in response to forces at the system boundaries^{1,2}, represents a fundamental goal of granular mechanics. The problem is of relevance to civil engineering, geophysics and physics^{3–5}, being important for the understanding of jamming, shear-induced yielding and mechanical response. Here we report measurements of the normal and tangential grain-scale forces inside a two-dimensional system of photoelastic disks that are subject to pure shear and isotropic compression. Various statistical measures show the underlying differences between these two stress states. These differences appear in the distributions of normal forces (which are more rounded for compression than shear), although not in the distributions of tangential forces (which are exponential in both cases). Sheared systems show anisotropy in the distributions of both the contact network and the contact forces. Anisotropy also occurs in the spatial correlations of forces, which provide a quantitative replacement for the idea of force chains. Sheared systems have long-range correlations in the direction of force chains, whereas isotropically compressed systems have short-range correlations regardless of the direction.

Under the action of external stresses, grains in dry granular materials form an inhomogeneous contact network, which carries most of the external load by way of force chains. The resultant network is different for shearing than for isotropic compression and is history-dependent owing to friction. Previous experiments^{6–8} have reported an exponential tail for the distribution of contact force magnitudes. This tail can be successfully predicted by many models^{9–11} with radically different mathematical structures and microscopic assumptions. Testing the validity of these models requires that the predicted force distributions be verified by measurements of full vectorial contact forces in the bulk of the sample. It is also important to find other distinguishing signatures characterizing the nature of force chain networks under different boundary conditions—an important goal of the present work.

In the following experiments, we visualize internal stresses in each grain and by solving the full inverse photoelastic problem^{12,13} for each disk, we obtain normal and tangential force components for each contact between disks. We use this microscopic contact force information to investigate differences in the distributions of contact forces, and the force chain structure, arising from two different types of loads: pure shear and isotropic compression. We find that forces have distinctive angular distributions and spatial correlations depending on the macroscopic preparation. In particular, forces have long-range correlations in the direction of force chains for sheared systems, but are correlated over a much shorter range, regardless of direction, for isotropically compressed systems.

Our experimental system is a two-dimensional (2D) array of approximately 2,500 bidisperse photoelastic (birefringent under

strain) disks subjected to pure shear and isotropic compression. Figure 1 shows a diagram of the experimental set-up and some typical images; details of the set-up and the experimental procedure are described in Fig. 1 legend. Although previous approaches^{14–17} have obtained contact forces using photoelastic techniques, they were neither automated nor suitable for a large enough number of

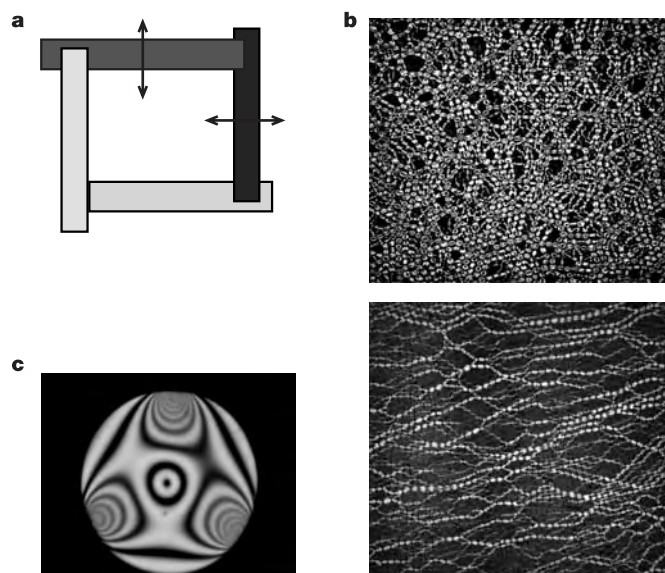


Figure 1 | Experimental set-up and representative data. **a**, Schematic diagram of the biaxial test cell. The biaxial test apparatus rests horizontally on a sheet of Plexiglas and is used to impart pure shear and isotropic compression. Motorized linear slides move two walls of the biaxial cell precisely and independently with a velocity of 0.024 cm s^{-1} . The system is illuminated from below and a high-resolution camera captures digital images from above. Each image captures roughly 250 particles located around the centre of the cell, roughly 10% of the total number of particles. The system is imaged through crossed circular polarizers. For each type of load, incremental deformations are applied in a quasi-static manner, beginning with a stress-free state. The sheared states are created by compressing in one direction and expanding by an equal amount in the other direction, with strains ($\epsilon_{xx} = \epsilon_{yy} = |\Delta L/L|$) ranging from 0 to 0.042. L ($\sim 40 \text{ cm}$) is the initial system length in the x or y direction. Isotropically compressed states are created by compressing in both directions with strains ranging from 0 to 0.016. The particles used in the experiment are either 0.8 cm or 0.9 cm in diameter and 0.6 cm in height, with a Young's modulus of 4 MPa and a friction coefficient of 0.8. The number ratio of small to large disks is 4:1. **b**, Typical system size images for an isotropically compressed state (top) and a sheared state (bottom). **c**, An example of the observed stress pattern for a single disk at the resolution ($\sim 0.01 \text{ cm}$ per pixel) used in these studies.

¹Department of Physics & Center for Nonlinear and Complex Systems, Duke University, Durham, North Carolina, 27708-0305, USA.

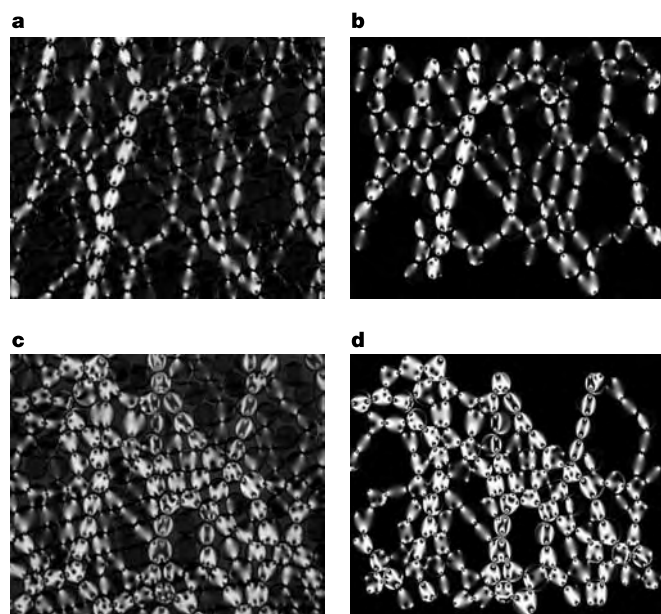


Figure 2 | Comparison of experimental images (a, c) and computed images (b, d). Top pair, a low-force sheared state. Bottom pair, a high-stress isotropically compressed state. Very weak forces are not resolved owing to a force threshold involved in the algorithm. Forces lower than 10% of the mean force are not computed.

particles to generate statistical information. Measurements of contact forces in our method are performed in a completely automated fashion for all disks except those at the boundary of the image.

Our algorithm for extracting contact forces (T.S.M. and R.P.B., manuscript in preparation) involves fitting the observed photoelastic pattern inside each disk to the plane elasticity solution¹⁸ for the stresses inside a disk, treating the force components as fitting parameters. In effect we solve the 2D isotropic elasticity equations for each disk, assuming a perfect line contact between three-dimensional (3D) cylinders. As long as the deformations are not too large, as is the case in our experiments, the 2D solution is a good approximation. As the experimental image is a highly nonlinear

function of the contact forces with the possibility of multiple solutions, we perform a nonlinear least-squares fit¹⁹ using about 500 data points for each disk. The best-fit parameter values are our measured contact forces.

For each realization, we compare the image computed using the best-fit contact forces and the original experimental image. Figure 2 shows two such realizations. The computed images (Fig. 2b, d) agree well with the experimental images (Fig. 2a, c) in terms of capturing the broad features of force chains. The error estimates found by calibration and by computing the average difference of pairs of forces at each contact are around 10% for low forces and 5% for mean-to-high forces. The agreement between the observed stress field inside each disk and the computed stress field validates the use of the 2D approximation. The data presented below (Figs 3 and 4) are obtained from five realizations for each stress state. We use the contact forces obtained by the procedure outlined above to investigate the distributions of contact forces and the anisotropy induced by external loads.

Figure 3 shows the distributions of the normal force, the tangential force and the ratio of tangential to normal force, for a sheared system and an isotropically compressed system. The normal and the tangential forces are normalized by the mean normal force. The normal force distribution for the sheared system (Fig. 3a) has a peak around the mean, a roughly exponential tail and a dip towards zero for forces lower than the mean. In contrast, for isotropically compressed systems, the normal force distribution (Fig. 3c) dips towards zero for forces below the mean, is broad around the mean, and decays faster for large forces compared to the sheared system. The tangential force distributions have a nearly exponential tail for forces larger than the mean for both the sheared (Fig. 3a) and the isotropically compressed system (Fig. 3c). The mean tangential forces are an order of magnitude smaller than the mean normal forces: a feature responsible for a smaller range of maximum tangential forces.

In order to investigate the role of friction in the system, we study the distribution of the variable $S = |F_t|/\mu F_n$, where μ is the static friction coefficient, F_t is the tangential force, and F_n is the normal force. The variable S gives information about how far away a contact is from the Coulomb failure criterion. If a contact is at the Coulomb failure criterion, $S = 1$. Figure 3b and d show the distributions of S for the sheared and the isotropically compressed system, respectively. For both types of loading conditions, the distribution of S shows that most of the contacts are much below the Coulomb failure condition.

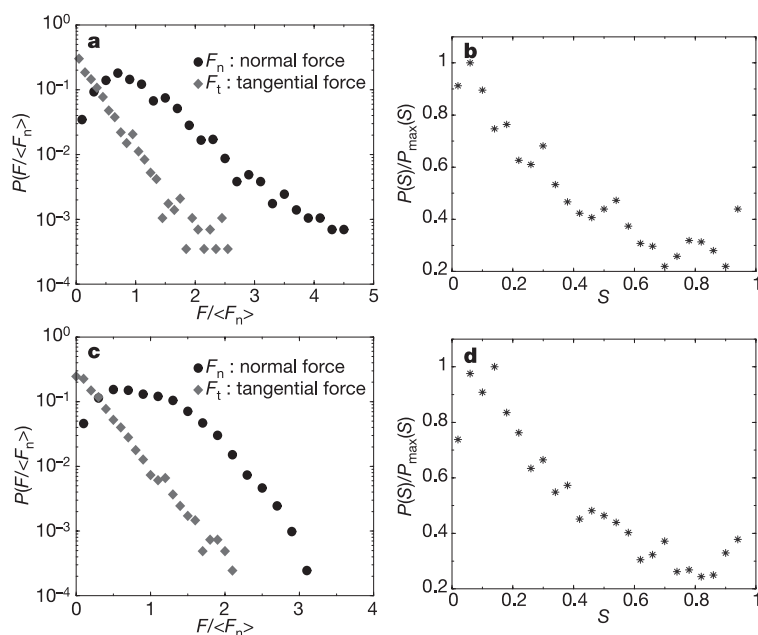


Figure 3 | Probability distributions of the normal forces, the tangential forces, and the mobilized friction, for the sheared and the isotropically compressed systems. The deformation ($\epsilon_{xx} = \epsilon_{yy} = |\Delta L/L|$) for isotropically compressed and sheared systems is 0.016 and 0.042, respectively. In **a** and **c**, the forces are normalized by the mean normal force, $\langle F_n \rangle$. **a**, Probability distributions of the normal (F_n) and the tangential (F_t) forces for the sheared system on a semi-log scale. **b**, Probability distribution of $S = |F_t|/\mu F_n$ for the sheared system, normalized by its maximum value, $P_{\max}(S)$. **c**, Probability distributions of the normal and the tangential forces for the isotropically compressed system on a semi-log scale. **d**, Probability distribution of $S = |F_t|/\mu F_n$ for the isotropically compressed system, normalized by its maximum value.

Numerical simulations of 3D granular systems in a silo geometry²⁰ report similar results for the distributions of the normal forces and the tangential forces in sheared systems. The distributions of the variable S (also called the mobilized friction) in the interior of the system reported in ref. 21 are qualitatively similar to our results for the sheared systems. The similarity in the distributions of the mobilized friction for the sheared and the isotropically compressed systems could be due to the fact that our sheared system is well below the shear yield limit and that our distributions are obtained from regions far away from the walls.

Our next focus of investigation is the shear-induced anisotropy. The anisotropy induced by an external load has two distinct effects; one, a purely geometric effect, is to introduce anisotropy in the contact network, and the other, a mechanical effect, is to develop an anisotropic force chain network and alter the stress distribution in the system. In order to investigate the geometric anisotropy, we study the distribution of contact angles for contacts carrying forces larger than the mean force. The sheared system (Fig. 4a) shows a strongly anisotropic distribution, with a large number of contacts aligned along the direction of the majority of force chains and a small number of contacts aligned in the direction perpendicular to it. In contrast, the isotropically compressed system exhibits a distribution with a six-fold symmetry (Fig. 4b), indicating that the contacts are distributed evenly along these directions. A simple measure characterizing the mechanical anisotropy is the angular variation of the mean normal force as shown in Fig. 4c and d, for the sheared and the isotropically compressed system, respectively. For the

sheared system, the mean normal force shows a periodic variation with peaks roughly in the direction of force chains. The variation can be adequately described by a second-order Fourier expansion (Fig. 4c), a result consistent with previous studies^{22,23}. In contrast, for the isotropically compressed system, the mean normal force is distributed randomly around an average value (Fig. 4d).

A more quantitative characterization of the force chain structure can be obtained by computing the 2D spatial correlation function of the magnitude of the forces on the particles. Here, the idea is to provide a well-defined quantitative measure to replace the rather vaguely defined notion of a force chain. Force correlation functions, in combination with force and angular distributions, provide an additional tool for discriminating among competing theories. Specifically, we computed $\langle F(\mathbf{x}) F(\mathbf{x} + \mathbf{r}) \rangle$, where $F(\mathbf{x})$ is the sum of the magnitudes of the contact forces on a particle, $\mathbf{x}$ gives the position vector of the point under consideration, and $\langle \dots \rangle$ implies an average over $\mathbf{x}$. Because we want to study the variation of this correlation in different directions, we do not average over angles. Thus, the correlation function gives spatial correlation between forces separated by a distance $\mathbf{r}$ in different directions. Figure 4e and f show the 2D correlation functions for the sheared system and the isotropically compressed system, respectively. The inset in each case shows a greyscale representation of the 2D correlation function. For the sheared system, the 2D correlation image (Fig. 4e, inset) reveals that the force correlations are much larger and of much longer range in the direction of the long force chains. Figure 4e, which depicts the variation of correlation function with distance in the

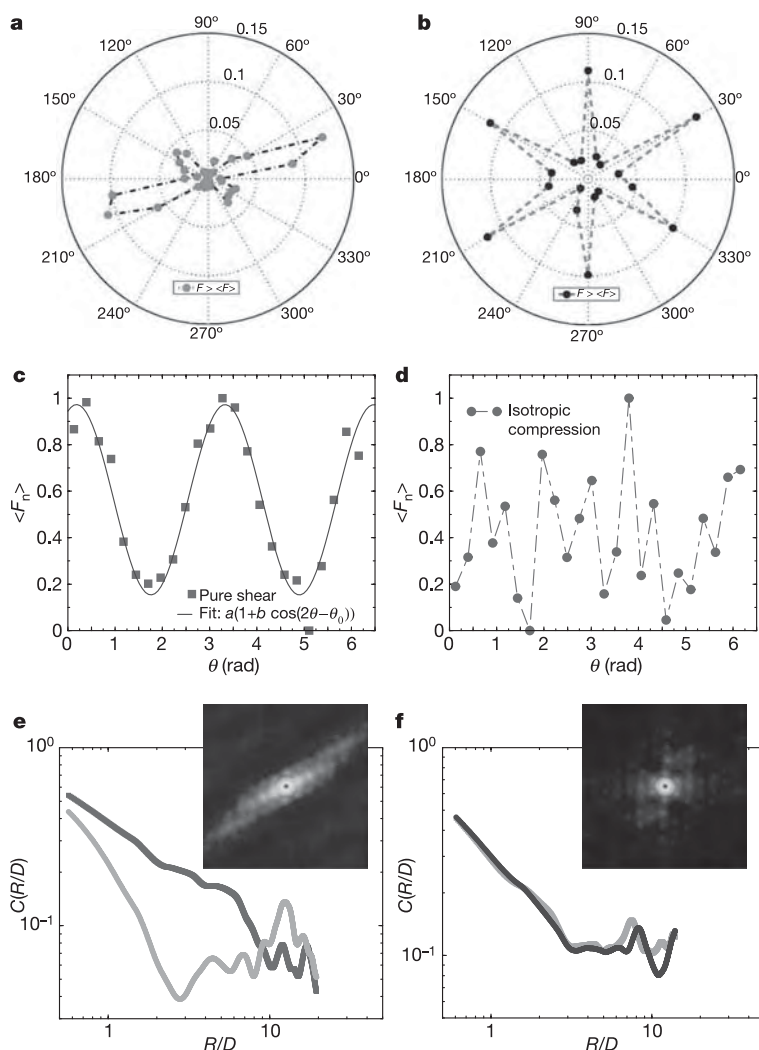


Figure 4 | Contact orientation, variation of the mean force and spatial correlations for the sheared and the isotropically compressed systems. The mean co-ordination numbers for the sheared system and the isotropically compressed system are 3.1 and 3.7, respectively. The deformations for each type of loads are the same as in Fig. 3. **a, b,** The distribution of contact angles of contacts carrying forces larger than the mean, for the sheared and the isotropically compressed system, respectively. **c, d,** The angular variation of the mean normal force, normalized by the maximum of the mean normal force (denoted only as $\langle F_n \rangle$ for clarity), for the sheared and the isotropically compressed system, respectively. Starting from the centre of the image, the image is divided into 24 angular bins of 15° each and the average normal force in each bin is plotted against the mean value of that bin. The angle (θ) in radians is measured with respect to the horizontal axis. The parameters of the fit in **c** are $a = 0.563$, $b = 0.727$ and $\theta_0 = 0.374$. The parameter a gives the mean force, b is a measure of anisotropy of the mean force, and θ_0 gives the direction in which the mean force is maximum. **e,** Spatial correlations for the sheared system in the direction of force chains and perpendicular to it. **f,** Spatial correlations for the isotropically compressed system in the same two directions used for the sheared system. The insets of **e** and **f** show the greyscale representations of the 2D correlations. The darker regions correspond to low correlation values, and brighter regions to high correlation values. The computations are performed in Fourier space, as the image sizes are large ($1,600 \times 1,152$). The plots are on a log-log scale with distance R normalized by the average diameter D of the disks.

direction of, and perpendicular to, the force chains, indicates that the spatial correlations in the force chain direction persist for up to ~ 15 particle diameters, whereas in the perpendicular direction they fall to background values within a couple of particle diameters. The correlation range in the direction of force chains is comparable to half the image size, which is 15 particle diameters in length. This length is also the Nyquist cut-off frequency of the Fourier method for computing the force correlation function in our computed images. The above data suggest the interesting possibility that the spatial force correlation may in fact be of a power-law type in the direction of long force chains for large system sizes.

In complete contrast to sheared systems, the isotropically compressed system has nearly uniform force correlation in all directions (Fig. 4f, inset). This behaviour is confirmed by examining the spatial correlation functions for the isotropically compressed system (Fig. 4f). For ease of comparison, the correlations are shown for the same two directions used for the sheared case. The correlation functions drop to background values in both directions after two particle diameters. Thus, no preferred direction is found in the isotropically compressed system, whereas in the sheared system the boundary stress creates an anisotropic stress state characterized by long force chains aligned in a specific direction. Our results strongly indicate that sheared systems exhibit not only geometric but also mechanical anisotropy, and that force correlations serve as an additional distinguishing signature characterizing stress-induced anisotropy in granular systems. The rather long range of correlations in the force chain direction is important for understanding the approach to continuum behaviour.

The observations reported here open up a new regime of comparison between theoretical models and experiments. One of the earliest computations of interparticle contact forces comes from numerical experiments¹¹, which are contact dynamics simulations of rigid, frictional particles under biaxial shear. Although the boundary conditions of the simulations and our experiments are not identical, these simulations seem to be the closest match to our experiments. As our current experimental resolution does not allow us to measure very small forces, we restrict our comparison to forces higher than the mean. In qualitative agreement with our data for the sheared systems, the simulations obtain an exponential tail for forces larger than the mean for the distributions of both the normal and the tangential forces.

Two recent models, which study the force distributions in 2D systems of frictionless particles under isotropic compression and shear, are also relevant for the present data. A force ensemble approach by Snoeijer *et al.*²³ assumes equal a priori probability for all force networks consistent with force balance constraints on each particle. In this model, the distribution of the magnitude of the force, $P(f)$, has a peak around the mean force, a finite value at $f = 0$, and a tail decaying faster than an exponential for compressed systems. For sheared systems, the model has an exponential regime for normal forces up to three times larger than the mean (J. Snoeijer, personal communication). A new lattice model²⁴ considers triangular lattices with force balance constraints on each node, which allows isotropic and anisotropic force chain networks. The lattice model prediction for the distribution of normal forces is an exponential tail for the sheared systems and a faster than exponential decay of the tail for compressed systems. The predictions of both models for sheared and isotropically compressed systems are in qualitative agreement with the present data. The absence of friction in both these models precludes a comparison of the distributions of tangential forces.

Looking towards the future, we are now in a position to address a variety of important issues, such as the nature of the jamming transition and the response function of a granular system. These issues are vital for gaining a deeper understanding of the macroscopic behaviour of granular systems from microscopic observations. Future goals will be the extension of these studies to yet larger systems and smaller forces.

Received 23 February; accepted 6 May 2005.

1. Geng, J. *et al.* Footprints in sand: The response of a granular material to local perturbations. *Phys. Rev. Lett.* **87**, 035506 (2001).
2. Reydellet, G. & Clément, E. Green's function probe of a static granular piling. *Phys. Rev. Lett.* **86**, 3308–3311 (2001).
3. Behringer, R. P. & Jenkins, J. T. (eds) *Powders and Grains 97* (Balkema, Rotterdam, 1997).
4. Nedderman, R. M. *Statics and Kinematics of Granular Materials* (Cambridge Univ. Press, Cambridge, UK, 1992).
5. Jaeger, H. M., Nagel, S. R. & Behringer, R. P. Granular solids, liquids, and gases. *Rev. Mod. Phys.* **68**, 1259–1273 (1996).
6. Meuth, D. M., Jaeger, H. M. & Nagel, S. R. Force distribution in a granular medium. *Phys. Rev. E* **57**, 3164–3169 (1998).
7. Løvøll, G., Måløy, K. J. & Flekkøy, E. G. Force measurements on static granular materials. *Phys. Rev. E* **60**, 5872–5878 (1999).
8. Tsoungui, O., Vallet, D. & Charmet, J. Use of contact area trace to study the force distributions inside 2D granular systems. *Granular Matter* **1**, 65–69 (1998).
9. Goldenberg, C. & Goldhirsch, I. Small and large scale granular statics. *Granular Matter* **6**, 87–97 (2003).
10. Snoeijer, J. H., van Hecke, M., Somfai, E. & van Saarloos, W. Force and weight distributions in granular media: Effects of contact geometry. *Phys. Rev. E* **67**, 030302(R) (2004).
11. Radjai, F., Wolf, D. E., Jean, M., Roux, S. & Moreau, J. in *Powders & Grains 97* (eds Behringer, R. P. & Jenkins, J. T.) 211–214 (Balkema, Rotterdam, 1997).
12. Frocht, M. M. *Photoelasticity* Vol. 1 (Wiley & Sons, New York, 1941).
13. Frocht, M. M. *Photoelasticity* Vol. 2 (Wiley & Sons, New York, 1948).
14. Drescher, A. & De-Josselin-de-Jong, G. Photoelastic verification of a mechanical model for the flow of a granular material. *J. Mech. Phys. Solids* **20**, 337–351 (1972).
15. Durelli, A. J. & Wu, D. Use of coefficients of influence to solve some inverse problems in plane elasticity. *J. Appl. Mech.* **50**, 288–296 (1983).
16. Shukla, A. & Nigam, H. A numerical-experimental analysis of the contact stress problem. *J. Strain Anal.* **20**, 241–245 (1985).
17. Paikowsky, S. G., DiRocco, K. J. & Xi, F. in *2nd Int. Conf. on Discrete Element Methods (DEM)* (eds Williams, J. R. & Mustoe, G. W.) 449–461 (IESL Publications, MIT, Boston, 1993).
18. Landau, L. *Theory of Elasticity* (Pergamon, New York, 1959).
19. Press, W. H., Teukolsky, H. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge, UK, 1992).
20. Landry, J. W., Grest, G. S., Silbert, L. E. & Plimpton, S. J. Confined granular packings: Structure, stress, and forces. *Phys. Rev. E* **67**, 041303 (2003).
21. Silbert, L. E., Grest, G. S. & Landry, J. W. Statistics of the contact network in frictional and frictionless granular packings. *Phys. Rev. E* **66**, 061303 (2002).
22. Cambou, B., Dubujet, Ph. & Nougier-Lehon, C. Anisotropy in granular materials at different scales. *Mech. Mater.* **36**, 1185–1194 (2004).
23. Snoeijer, J. H., Vlucht, T. J. H., van Hecke, M. & van Saarloos, W. Force network ensemble: A new approach to static granular matter. *Phys. Rev. Lett.* **92**, 054302 (2004).
24. Tighe, B. P., Socolar, J. E. S., Schaeffer, D. G., Mitchener, W. G. & Huber, M. L. Force distributions in a triangular lattice of rigid bars. Preprint at (<http://arXiv.org/cond-mat/0505003>) (2005).

Acknowledgements This work was supported by NSF DMR, NSF DMS, and NASA. We thank J. Snoeijer and collaborators, J. Socolar and B. Tighe for providing their data and for discussions. We thank H. Phatnani for critical reading of the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.P.B. (bob@phy.duke.edu).

Astronomical pacing of late Palaeocene to early Eocene global warming events

Lucas J. Lourens¹, Appy Sluijs², Dick Kroon³, James C. Zachos⁴, Ellen Thomas⁵, Ursula Röhl⁶, Julie Bowles⁷ & Isabella Raffi⁸

At the boundary between the Palaeocene and Eocene epochs, about 55 million years ago, the Earth experienced a strong global warming event, the Palaeocene–Eocene thermal maximum^{1–4}. The leading hypothesis to explain the extreme greenhouse conditions prevalent during this period is the dissociation of 1,400 to 2,800 gigatonnes of methane from ocean clathrates^{5,6}, resulting in a large negative carbon isotope excursion and severe carbonate dissolution in marine sediments. Possible triggering mechanisms for this event include crossing a threshold temperature as the Earth warmed gradually⁷, comet impact⁸, explosive volcanism^{9,10} or ocean current reorganization and erosion at continental slopes¹¹, whereas orbital forcing has been excluded¹². Here we report a distinct carbonate-poor red clay layer in deep-sea cores from Walvis ridge¹³, which we term the Elmo horizon. Using orbital tuning, we estimate deposition of the Elmo horizon at about 2 million years after the Palaeocene–Eocene thermal maximum. The Elmo horizon has similar geochemical and biotic characteristics as the Palaeocene–Eocene thermal maximum, but of smaller magnitude. It is coincident with carbon isotope depletion events in other ocean basins, suggesting that it represents a second global thermal maximum. We show that both events correspond to maxima in the ~405-kyr and ~100-kyr eccentricity cycles that post-date prolonged minima in the 2.25-Myr eccentricity cycle, implying that they are indeed astronomically paced.

Biotic phenomena similar to those characterizing the Palaeocene–Eocene thermal maximum (PETM) have been locally recorded in the upper Palaeocene to lower Eocene, indicating the possibility of additional hyperthermal events, though of smaller magnitude^{14–16}. Several short, negative carbon isotope shifts of up to 1‰ at deep-sea sites resemble the much larger-amplitude carbon isotope excursion at the PETM¹². Orbital tuning suggested that these transients were controlled by maxima in the short-term eccentricity cycles, whereas the PETM carbon isotope excursion allegedly occurred near a minimum in the ~405-kyr eccentricity cycle, excluding orbital forcing as a triggering mechanism for the latter¹².

One objective of Ocean Drilling Program (ODP) Leg 208 on the Walvis ridge (subtropical southeastern Atlantic Ocean) was to search for hyperthermal events within the lower Cenozoic greenhouse climate record. We recovered continuous, undisturbed lower Palaeogene successions at five sites along a 2-km water depth transect in multiple (mostly advanced piston core) holes. This resulted in the first complete early Palaeogene deep-sea record accumulated at

relatively high sedimentation rates¹³. The uppermost Palaeocene and lower Eocene are composed of foraminifer-bearing nannofossil ooze, with a few chert layers and two deep-red clay layers marking the PETM and a younger distinctive horizon, named Elmo. Magneto-stratigraphic results on Site 1262 (see Supplementary Information) reveal that the Elmo horizon at 117.1–117.2 m composite depth (m.c.d.) is slightly older than the chron C24r/C24n reversal boundary (115–116 m.c.d.) (Supplementary Fig. 1) and occurs within the lower part of NP11.

The Elmo horizon is 10–15 cm thick, and characterized by elevated magnetic susceptibility (MS) values at all sites (Fig. 1). Analysis of the CaCO₃ content (expressed in weight per cent: wt%) of the deepest Site 1262 (palaeodepth¹³, 3,600 m), intermediate Site 1266 (palaeodepth, 2,600 m) and shallowest Site 1263 (palaeodepth, 1,500 m) reveal that the increase in MS is linearly related to a drop in CaCO₃ wt% (Supplementary Fig. 2). The CaCO₃ wt% declines from 90–95 below, to ~40 within, the red clay. High-resolution bulk carbon isotope records ($\delta^{13}\text{C}_{\text{bulk}}$) of Sites 1262, 1265, 1266 and 1267 reveal a negative excursion of 1.0–1.2‰ from below the first decline in CaCO₃ wt% into the Elmo (Fig. 1). The $\delta^{13}\text{C}_{\text{bulk}}$ of Site 1263 shows the largest depletion (1.4–1.6‰), suggesting that the red clay layer at this site with the highest sedimentation rate¹³ is stratigraphically the most complete and/or least affected by the dissolution of primary calcite and the presence of reworked or secondary calcite. The post-Elmo interval mirrors the typical PETM signature with an exponential recovery to pre-excursion $\delta^{13}\text{C}_{\text{bulk}}$ values. The bulk carbonate oxygen isotope record ($\delta^{18}\text{O}_{\text{bulk}}$) of Site 1263 shows a negative excursion of ~1.6‰ (Fig. 2).

From Site 1263, we analysed the stable isotopic composition of individual specimens (>300 μm size fraction) of the surface-dwelling planktonic foraminifer *Acarinina soldadoensis* and the benthic foraminifers *Cibicides* spp. and *Anomalinoidea* spp. (Fig. 2). The planktonic foraminiferal data show a much larger inter-specimen variability within each sample (especially within the Elmo horizon) than the benthic data. The (smoothed) carbon isotope record of *A. soldadoensis* ($\delta^{13}\text{C}_{A.\text{soldadoensis}}$) resembles the pattern of the bulk record, but shows a significantly larger negative excursion (~2.5‰). The carbon isotope shift is much smaller in the benthic foraminiferal record than in *A. soldadoensis*, but the (~1‰) trend through the carbonate-rich intervals equals that of the bulk and planktonic isotope records. Benthic foraminifera species richness is low and assemblages are dominated by diminutive *Nuttallides truempyi* and *Abyssamina* spp. species in the Elmo horizon. The

¹Faculty of Geosciences, Department of Earth Sciences, and ²Laboratory of Palaeobotany and Palynology, Department of Palaeoecology, Utrecht University, Budapestlaan 4, 3584 CD Utrecht, The Netherlands. ³Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands. ⁴Earth Science Department, University of California, Santa Cruz, Earth and Marine Sciences Building, Santa Cruz, California 95064, USA. ⁵Department of Earth and Environmental Sciences, Wesleyan University, 265 Church Street, Middletown, Connecticut 06459-0139, USA, and Center for the Study of Global Change, Department of Geology and Geophysics, Yale University, PO Box 208109, New Haven, Connecticut 06520-8109, USA. ⁶DFG Research Center for Ocean Margins (RCOM), University of Bremen, Leobener Strasse, 28359 Bremen, Germany. ⁷Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, MC0208, La Jolla, California 92093, USA. ⁸Facoltà di Scienze, Dipartimento Scienze della Terra, Università "G. d'Annunzio" di Chieti, Campus Universitario Madonna delle Piane, Via dei Vestini 31, 66013 Chieti Scalo, Italy.

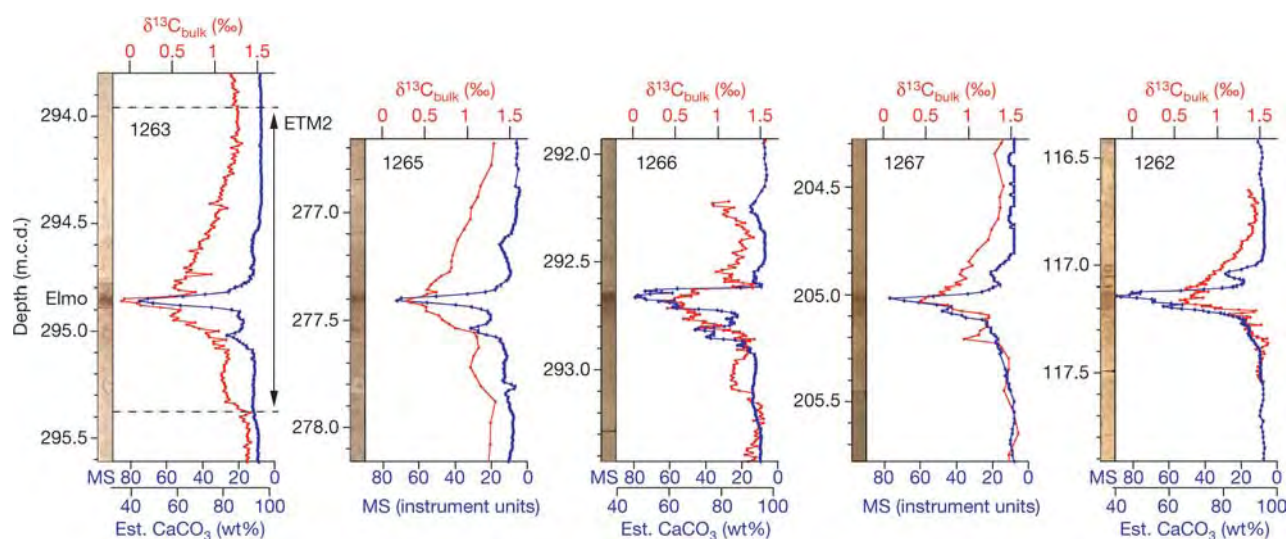


Figure 1 | Bulk carbonate $\delta^{13}\text{C}$ and magnetic susceptibility (MS) records across the Elmo horizon at five ODP Leg 208 sites. The CaCO_3 wt% axes are estimates based on linear correlation with MS measurements on the same samples (Methods). Site numbers are given at the top left of each panel.

Site numbers and water depths (m) are as follows: 1263, 2,717 m; 1265, 3,060 m; 1266, 3,798 m; 1267, 4,355 m; 1262, 4,755 m. Digital images of the lithology are plotted at the left site of each panel.

few measured benthic isotope values of the Elmo horizon, representing *Anomalinoidea* spp. specimens (*Cibicidoides* spp. >300 μm are absent in the Elmo horizon), are similar to those from outside the clay layer, indicating that these are probably derived from bioturbated specimens. This suggests that large-sized benthic foraminifera were absent during deposition of the Elmo horizon, as commonly observed for the PETM¹⁷. The presence of light-coloured burrows within the red clay layer documents bioturbation, and could explain the scatter in the planktonic isotope values, and the less strong $\delta^{13}\text{C}_{\text{bulk}}$ excursion relative to $\delta^{13}\text{C}_{A.\text{soldadoensis}}$. This possibility does not rule out that the magnitude of the excursion in the deep sea could have been damped owing to the larger carbon mass of this reservoir. The maximum oxygen isotope shift in *A. soldadoensis* ($\delta^{18}\text{O}_{A.\text{soldadoensis}}$) across the Elmo horizon is comparable to that of the $\delta^{18}\text{O}_{\text{bulk}}$ record. There is only a $\sim 0.6\text{‰}$ shift in the benthic

oxygen isotope record, either because no *in situ* large benthic foraminifera are present in the Elmo horizon or changes in bottom water temperatures were minor.

To unravel the orbital relationship between the Elmo horizon and PETM, we studied the cyclic sedimentary patterns of the interlying interval in continuous spliced cores derived from advanced piston core holes only. Spectral analysis was applied on the colour reflectance (L^*) of Site 1267, and L^* and MS of Site 1262 (see Supplementary Information). The spectra of all records revealed the dominance of the long ($\sim 405\text{-kyr}$) and short ($\sim 100\text{-kyr}$) eccentricity cycles (Supplementary Fig. 3). Both components were extracted and could be unambiguously correlated between the records of these sites (Fig. 3). Four long-term maxima in MS (minima in L^*) occur between the Elmo horizon and PETM. The Elmo horizon corresponds to a fifth long-term MS maximum (L^* minimum) and a

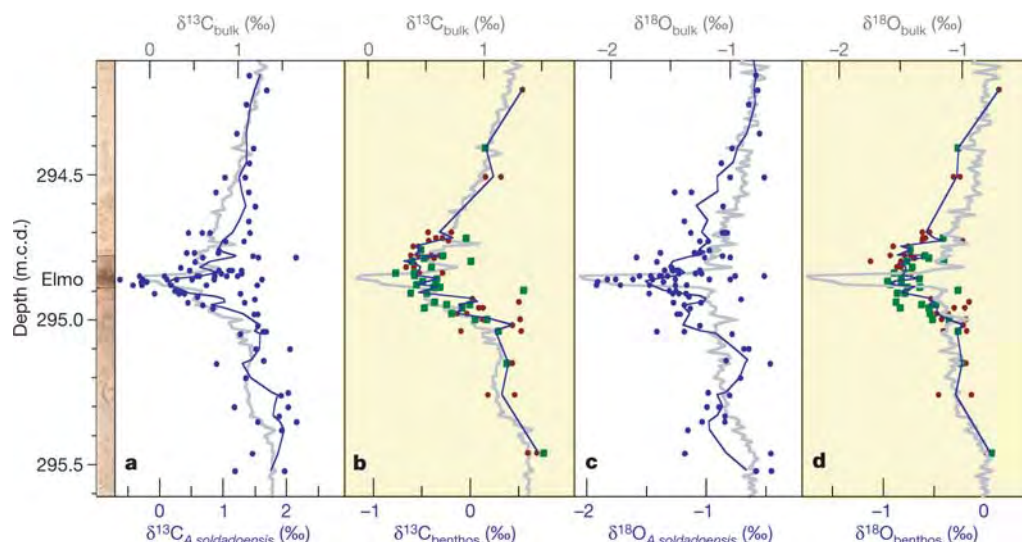


Figure 2 | Stable isotope series of bulk sediment and single foraminifer specimens across the Elmo horizon at Site 1263. **a**, The $\delta^{13}\text{C}$ values (blue dots) of the surface dwelling planktonic foraminifer *A. soldadoensis* ($\delta^{13}\text{C}_{A.\text{soldadoensis}}$). **b**, The $\delta^{13}\text{C}$ values of the bottom dwelling benthic foraminifers ($\delta^{13}\text{C}_{\text{benthos}}$) *Cibicidoides* spp. (red squares) and *Anomalinoidea* spp. (green dots). **c** and **d** as in **a** and **b** but for $\delta^{18}\text{O}$. Grey lines in **a** and **b**, and in **c** and **d**, indicate respectively the $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ values of the bulk sediment ($\delta^{13}\text{C}_{\text{bulk}}$, $\delta^{18}\text{O}_{\text{bulk}}$). Blue lines represent three-point moving averages on averaged values of duplicate analyses of a sample.

spp. (green dots). **c** and **d** as in **a** and **b** but for $\delta^{18}\text{O}$. Grey lines in **a** and **b**, and in **c** and **d**, indicate respectively the $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ values of the bulk sediment ($\delta^{13}\text{C}_{\text{bulk}}$, $\delta^{18}\text{O}_{\text{bulk}}$). Blue lines represent three-point moving averages on averaged values of duplicate analyses of a sample.

short-term MS maximum (L^* minimum) cycle. The red clay layer associated with the PETM ends in a long-term MS minimum (L^* maximum). If there were 11 climate precession cycles in the PETM interval¹⁸, then its carbon isotope excursion corresponds to a maximum (minimum) in the long-term MS (L^*) cycle, similar to the Elmo.

A definite tuning of the early Eocene to astronomical computations is complicated, because the precision of the orbital solution more than 45 Myr ago is limited^{19,20}. Tuning is in principle possible for the 405-kyr eccentricity cycle, because of its longer duration of stability^{19,20} (Fig. 3). At 50 Myr ago, the absolute uncertainty in time is about 20 kyr (ref. 20), but this did not lead to an astronomically

tuned timescale owing to large uncertainties in radiometric age constraints for this time interval²¹. A second uncertainty derives from the chaotic behaviour of the inner planets related to the resonant argument $\theta = (s_4 - s_3) - 2(g_4 - g_3)$, where g_3 , g_4 are related to the precession of the perihelion and s_3 , s_4 to the precession of the node of Earth and Mars²⁰. This causes a large uncertainty in the determination of the time when the relatively stable ~ 2.4 -Myr beat in eccentricity evolved from the ~ 1.2 -Myr period when $(s_4 - s_3) - (g_4 - g_3) = 0$ (that is, ~ 2.25 Myr in the nominal La2004²⁰ and R7¹⁹ solutions between 53–57 Myr ago). This problem limits an accurate age determination of successive minima in this very long eccentricity cycle and the related intervals of reduced amplitude

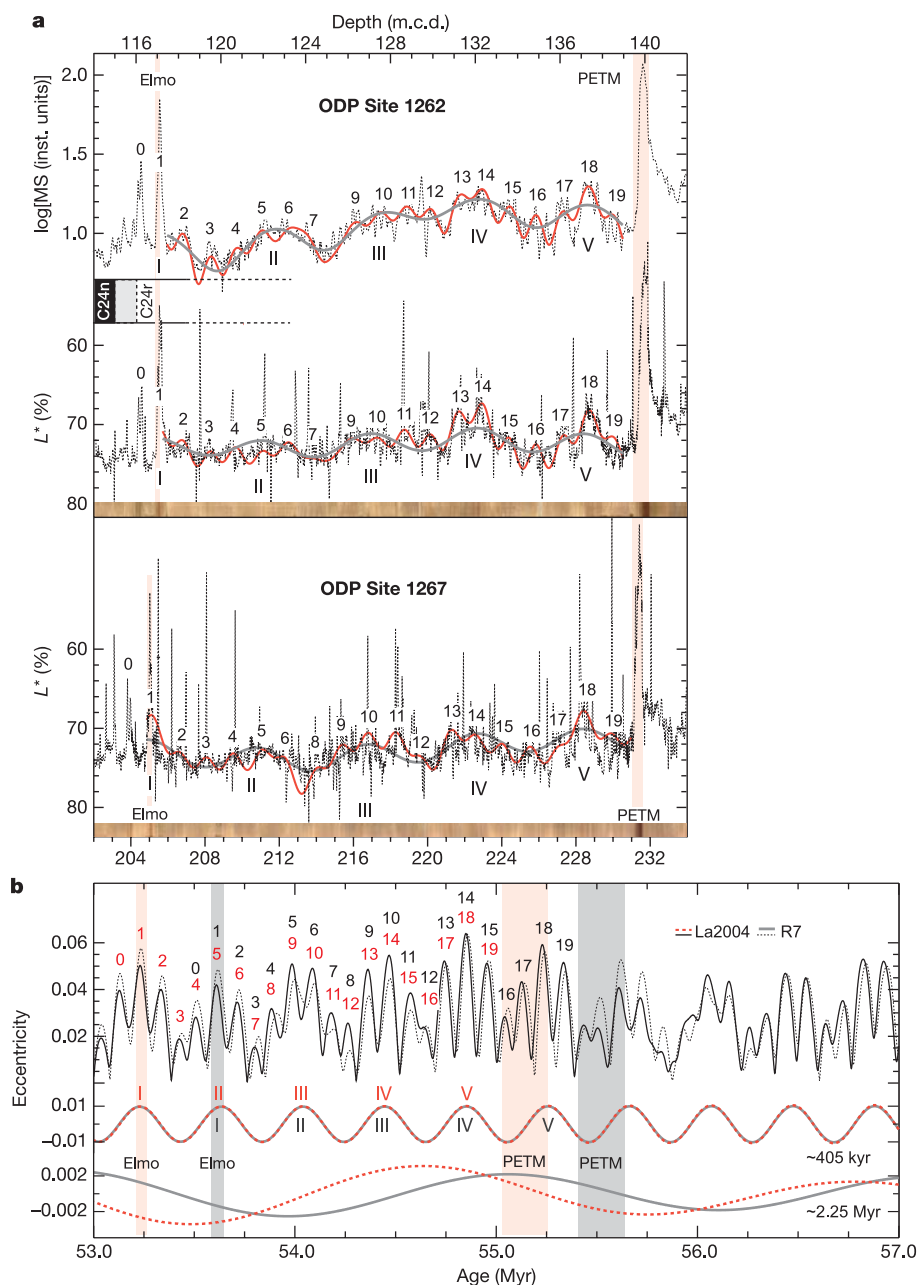


Figure 3 | Astronomical tuning of the lower Eocene sediments at Walvis ridge to two different orbital computations. **a**, The extracted short (red lines) and long (grey lines) eccentricity-related cycles from magnetic susceptibility (MS) and colour reflectance (L^*) of sites 1262 and 1267 (dashed lines) represent respectively the 97.5% and 99.5% significant peaks in the CLEAN spectra (Supplementary Fig. 4). **b**, Correlation of the long-

term (I–V) and short-term (0–19) eccentricity-related maxima to their correlative cycles in the R7²³ (grey) and La2004²⁴ (red) orbital solutions. The 405-kyr and ~ 2.25 -Myr cycles of eccentricity were extracted using a gaussian filter with a frequency of 0.00247 ± 0.0001 (kyr^{-1}) and of 0.00043 ± 0.0001 (kyr^{-1}), respectively.

changes in the short eccentricity cycle, and explains the offset between the ~ 2.25 -Myr cycles of the nominal La2004 and R7 solutions in the studied time interval (Fig. 3).

Because of these uncertainties, only a floating tuning could be realized (see also Supplementary Information). First, we emphasize that MS maxima (L^* minima) correlate to eccentricity maxima based on the distinct amplification of the precession-related lithological changes during the long- and short-term MS maxima (Supplementary Fig. 4). This observation is crucial, because it implies that the carbon isotope shifts associated with the PETM and Elmo horizon also correspond to maxima in the long and short eccentricity cycles (Fig. 3). Second, we correlated the (on average less amplified) ~ 100 -kyr cycles within the second (II; Fig. 3) ~ 405 -kyr cycle to the minimum in the ~ 2.25 -Myr cycle at 53.5 Myr ago (La2004) or 54.0 Myr ago (R7) (Supplementary Fig. 5). Using this first order calibration, we tuned all long and short eccentricity cycles, implying that all cycles should be shifted one ~ 405 -kyr cycle older in R7 than in the nominal La2004 solution (Fig. 3). As a result, the Elmo horizon correlates with the short eccentricity maximum at ~ 53.235 Myr ago (La2004) or ~ 53.620 Myr ago (R7), and the onset of the PETM carbon isotope excursion correlates with the long eccentricity maximum centred at ~ 55.270 Myr ago (La2004) or ~ 55.675 Myr ago (R7). This tuning implies that both events occurred briefly after a period of low-amplitude, short eccentricity changes associated with a minimum in the very long-term orbital perturbation of ~ 2.25 Myr.

The $\delta^{13}\text{C}_{\text{bulk}}$ negative shift of 1.4–1.6‰ in the Elmo horizon at Site 1263 is, with exception of the PETM carbon isotope excursion, of an unusually large magnitude for the early Palaeogene¹². Its position just below C24n/C24r and within NP11 suggests that this excursion correlates to the $\sim 1\%$ depletion characterizing the H1 event in the North Atlantic (DSDP Site 550 and ODP Site 1051), the Southern Ocean (ODP Site 690) and the Pacific Ocean (DSDP Site 577)¹² (see Supplementary Information). Moreover, the palaeosol carbonate isotope record from the Bighorn basin also shows a strong negative $\delta^{13}\text{C}$ excursion just below C24n/C24r²², indicating that the carbon isotope excursion is global and recorded in both marine and terrestrial basins (Supplementary Fig. 6). The Elmo horizon, however, has yet not been recognized at other locations, although the H1 event is accompanied by high MS values at Sites 550 and 690¹². Hence, the large drop in CaCO_3 wt% in the Walvis ridge cores probably indicates a major global ocean lysocline shoaling, but in contrast to the PETM²³, the calcite compensation depth appears to have remained below the palaeodepth of Site 1262. Application of the empirical temperature– $\delta^{18}\text{O}$ relationship^{24,25} indicates furthermore that the $\sim 1\%$ $\delta^{18}\text{O}_{\text{A.soldadoensis}}$ change within the Elmo horizon reflects a sea surface temperature rise of at least 3–4 °C, about half of the mid- to high-latitude sea surface temperature changes estimated for the PETM²⁶. All this suggests that the Elmo horizon characterizes a second pronounced early Eocene thermal maximum (ETM2; Fig. 1), similar to the PETM in both orbital and biogeochemical aspects, but of approximately half its amplitude in carbon isotope excursion, rise in sea surface temperature, and carbonate dissolution.

The linkage of both events to a similar orbital configuration disagrees with Cramer *et al.*¹², who related the PETM to a minimum and H1 (ETM2 equivalent) to a maximum in a ~ 405 -kyr cycle, thereby promoting the comet impact hypothesis⁸. In addition, according to their tuning the interval between the PETM carbon isotope excursion and C24n/C24r should span ~ 1.5 Myr, which is significantly shorter than the five ~ 405 -kyr cycles (~ 2 Myr) that we found. These discrepancies can probably be attributed to large uncertainties in their approach of using low-amplitude bulk carbon isotope transient excursions and counts of poorly expressed lithological cycles from incomplete successions¹². Hence, we suggest that the extreme seasonal contrast at both hemispheres during eccentricity maxima increased intermediate seawater temperatures,

thereby triggering the release of oceanic methane hydrates. In this respect, the critical conjunction of short, long and very long eccentricity cycles and the long-term late Palaeocene to early Eocene warming trend may have favoured the build-up of a significant methane hydrate reservoir before its release during both events, thereby excluding unique mechanisms for explaining the PETM^{5–11,27}. The less extreme signal of ETM2 may reflect the inability of the methane hydrate reservoir to return to pre-PETM dimensions, especially under the warm conditions that prevailed in the interval spanning the two events¹. Above ETM2 (H1) an increasing number of low-amplitude carbon transients occurred, of which the first, H2¹², seems to correspond with the two thin brown layers one 100-kyr cycle above the Elmo horizon (number 0 in Fig. 3), suggesting that the threshold for dissociation of clathrates was low during the early Eocene climatic optimum¹, enabling even the short eccentricity cycles to trigger minor methane releases.

METHODS

Sampling and CaCO_3 wt% analyses. Discrete sediment samples were collected at a 0.5–1-cm spacing across the Elmo horizon in holes 1262A, 1263C and 1266C. All samples were freeze-dried and analysed for magnetic susceptibility per gram sediment (MS g^{-1}) using an AGICO KLY-3 device. These records were compared to the split core point magnetic susceptibility (PMS) and whole core MS of the multiple sensor track (MS-MST) measurements obtained during Leg 208¹³. We converted all MS data to the MS-MST scale by performing linear regression analyses between MS g^{-1} and PMS (Supplementary Fig. 2) and the conversion of PMS to MS-MST using the equation $\text{MST} = (\text{PMS} \times 2.0683) + 7.8257$ ($R^2 = 0.99$)¹³.

Every fourth sample (but all within the Elmo horizon) was used for calcium carbonate analyses. The CaCO_3 wt% was based on the amount of total carbon combusted with the Fison NA 1500 CNS analyser. Analytical precision and accuracy were determined by comparison with an international standard (BCR-71) and in-house standards (F-TURB, MM-91). The relative standard deviations, analytical precision and accuracy were better than 3%. Several samples prepared for palynological studies revealed that no significant amount of organic carbon was present, with an uncertainty smaller than the analytical precision. A regression analysis between the CaCO_3 wt% and the MS g^{-1} (converted to the MS-MST scale) was applied (Supplementary Fig. 2) to obtain the estimated CaCO_3 wt% scale (Fig. 1).

Stable isotopes. Bulk stable isotope measurements were carried out for all sites with an average spacing of 4 cm, but in 0.5–1-cm resolution through the Elmo horizon. The isotope measurements were carried out using an ISOCARB common bath carbonate preparation device linked on-line to a VG SIRA24 mass spectrometer. Isotope values were calibrated to the Pee Dee Belemnite (PDB) scale. Analytical precision was determined by replicate analyses and by comparison to international (IAEA-CO1 and NBS19) and in-house (NAXOS) carbonate standards, showing standard deviations of $<0.06\%$ and $<0.1\%$ for $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$, respectively.

Stable isotope measurements of individual planktonic and benthic foraminiferal specimens were carried out using a CARBO-KIEL automated carbonate preparation device linked on-line to a Finnigan MAT252 mass spectrometer. Specimens were hand picked from the $>300\ \mu\text{m}$ fraction and cleaned in ethanol in an ultrasonic bath for 30 s. Calibration to the international carbonate standard NBS19 revealed an analytical precision better than 0.03‰ and 0.05‰ for $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$, respectively.

Received 17 November 2004; accepted 10 May 2005.

Published online 8 June 2005.

1. Zachos, J. C., Pagani, M., Sloan, L. C., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
2. Norris, R. D. & Röhl, U. Carbon cycling and chronology of climate warming during the Palaeocene/Eocene transition. *Nature* **401**, 775–778 (1999).
3. Kennett, J. P. & Stott, L. D. Abrupt deep-sea warming, palaeoclimatographic changes and benthic extinctions at the end of the Palaeocene. *Nature* **353**, 225–229 (1991).
4. Koch, P. L., Zachos, J. C. & Gingerich, P. D. Correlation between isotope records in marine and continental carbon reservoirs near the Palaeocene/Eocene boundary. *Nature* **358**, 319–322 (1992).
5. Dickens, G. R., O'Neil, J. R., Rea, D. K. & Owen, R. M. Dissociation of oceanic methane hydrate as a cause of the carbon isotope excursion at the end of the Paleocene. *Paleoceanography* **10**, 965–971 (1995).
6. Dickens, G. R., Castillo, M. M. & Walker, J. C. G. A blast of gas in the latest

- Paleocene: Simulating first-order effects of massive dissociation of oceanic methane hydrate. *Geology* **25**, 259–262 (1997).
7. Thomas, E. & Shackleton, N. J. in *Correlation of the Early Paleogene in Northwestern Europe* (eds Knox, R. W. O. B., Corfield, R. M. & Dunay, R. E.) 401–441 (Special Publication 101, Geological Society, London, 1996).
 8. Kent, D. V. *et al.* A case for a comet impact trigger for the Paleocene/Eocene thermal maximum and carbon isotope excursion. *Earth Planet. Sci. Lett.* **211**, 13–26 (2003).
 9. Bralower, T. J. *et al.* High-resolution records of the late Paleocene thermal maximum and circum-Caribbean volcanism: Is there a causal link? *Geology* **25**, 963–966 (1997).
 10. Schmitz, B. *et al.* Basaltic explosive volcanism, but no comet impact, at the Paleocene-Eocene boundary: high-resolution chemical and isotopic records from Egypt, Spain and Denmark. *Earth Planet. Sci. Lett.* **225**, 1–17 (2004).
 11. Katz, M. E., Cramer, B. S., Mountain, G. S., Katz, S. & Miller, K. G. Uncorking the bottle: What triggered the Paleocene/Eocene thermal maximum methane release? *Paleoceanography* **16**, 1–14 (2001).
 12. Cramer, B. S., Wright, J. D., Kent, D. V. & Aubry, M.-P. Orbital climate forcing of $\delta^{13}\text{C}$ excursions in the late Paleocene–early Eocene (chons C24n–C25n). *Paleoceanography* **18**, doi:10.1029/2003PA000909 (2003).
 13. Zachos, J. C., *et al.* in *Early Cenozoic Extreme Climates: The Walvis Ridge Transect* (eds Zachos, J. C., Kroon, D. & Blum, P.) (Ocean Drilling Program, College Station, Texas, 2004).
 14. Thomas, E. & Zachos, J. C. Was the late Paleocene thermal maximum a unique event? *Geol. Föhr. Stockh. Föhr. [Trans. Geol. Soc. Stockholm]* **122**, 169–170 (2000).
 15. Bujak, J. P. & Brinkhuis, H. in *Late Paleocene - Early Eocene Biotic and Climatic Events in the Marine and Terrestrial Records* (eds Aubry, M.-P., Lucas, S. G. & Berggren, W. A.) 277–295 (Columbia Univ. Press, New York, 1998).
 16. Röhl, U., Norris, R. D. & Ogg, J. G. in *Causes and Consequences of Globally Warm Climates in the Early Paleogene* (eds Wing, S. L., Gingerich, P. D., Schmitz, B. & Thomas, E.) 567–589 (Special Paper 369, Geological Society of America, Boulder, Colorado, 2003).
 17. Thomas, E., Zachos, J. C. & Bralower, T. J. in *Warm Climates in Earth History* (eds Huber, B. T., MacLeod, K. & Wing, S. L.) 132–160 (Cambridge Univ. Press, Cambridge, 2000).
 18. Röhl, U., Bralower, T. J., Norris, G. & Wefer, G. New chronology for the late Paleocene thermal maximum and its environmental implications. *Geology* **28**, 927–930 (2000).
 19. Varadi, F., Bunnegar, B. & Ghil, M. Successive refinements in long-term integrations of planetary orbits. *Astrophys. J.* **592**, 620–630 (2003).
 20. Laskar, J. *et al.* A long term numerical solution for the insolation quantities of Earth. *Astron. Astrophys.* **428**, 261–285 (2004).
 21. Machlus, M., Hemming, S. R., Olsen, P. E. & Christie-Blick, N. Eocene calibration of geomagnetic polarity time scale reevaluated: Evidence from the Green River Formation of Wyoming. *Geology* **32**, 137–140 (2004).
 22. Koch, P. L. *et al.* in *Causes and Consequences of Globally Warm Climates in the Early Paleogene* (eds Wing, S. L., Gingerich, P. D., Schmitz, B. & Thomas, E.) 49–64 (Special Paper 369, Geological Society of America, Boulder, Colorado, 2003).
 23. Zachos, J. C. *et al.* Rapid acidification of the ocean during the Paleocene-Eocene Thermal Maximum. *Science* (in the press).
 24. O'Neil, J. R., Clayton, R. N. & Mayeda, T. K. Oxygen isotope fractionation in divalent metal carbonates. *J. Chem. Phys.* **51**, 5547–5558 (1969).
 25. Shackleton, N. J. Oxygen isotope analyses and Pleistocene temperatures reassessed. *Nature* **215**, 15–17 (1967).
 26. Zachos, J. C. *et al.* A transient rise in tropical sea surface temperature during the Paleocene-Eocene thermal maximum. *Science* **302**, 1151–1154 (2003).
 27. Svensen, H. *et al.* Release of methane from a volcanic basin as a mechanism for initial Eocene global warming. *Nature* **429**, 542–545 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research used samples and data provided by the Ocean Drilling Program (ODP). This work was supported by the Netherlands Organisation for Scientific Research (L.J.L., A.S. and D.K.), Utrecht Biogeology Centre (A.S.), Deutsche Forschungsgemeinschaft (U.R.), and the National Science Foundation (J.C.Z., E.T. and J.B.). We thank the scientific and non-scientific crew of ODP Leg 208, J. Suhonen in particular, and G. Ittman, A. E. van Dijk, G. M. Ganssen, S. J. A. Jung, H. B. Vonhof, P. L. Koch, H. Brinkhuis, F. J. Hilgen, T. Kouwenhoven and J. W. Zachariasse for technical support, advice and comments.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.J.L. (llorens@geo.uu.nl).

LETTERS

Space geodetic evidence for rapid strain rates in the New Madrid seismic zone of central USA

R. Smalley Jr^{1,2}, M. A. Ellis^{1,2}, J. Paul¹ & R. B. Van Arsdale²

In the winter of 1811–1812, near the town of New Madrid in the central United States and more than 2,000 km from the nearest plate boundary, three earthquakes within three months shook the entire eastern half of the country and liquefied the ground over distances far greater than any historic earthquake in North America^{1,2}. The origin and modern significance of these earthquakes, however, is highly contentious³. Geological evidence demonstrates that liquefaction due to strong ground shaking, similar in scale to that generated by the New Madrid earthquakes, has occurred at least three and possibly four times in the past 2,000 years (refs 4–6), consistent with recurrence statistics derived from regional seismicity⁷. Here we show direct evidence for rapid strain rates in the area determined from a continuously operated global positioning system (GPS) network. Rates of strain are of the order of 10^{-7} per year, comparable in magnitude to those across active plate boundaries, and are consistent with known active faults within the region. These results have significant implications for the definition of seismic hazard and for processes that drive intraplate seismicity.

Current models for generating crustal earthquakes require a means of generating and replenishing strain energy in the Earth's crust, a process that readily occurs along the boundaries of rigid tectonic plates⁸. Large, frequent earthquakes therefore require rapid accumulation in the crust of a significant amount of strain energy. Such strain accumulation is typically observed as differential velocities measured at the Earth's surface via space geodetic surveys. Before the establishment of the permanent GPS array in mid-America (GAMA), space-based geodesy had failed to yield significant differential surface velocities in the New Madrid seismic zone^{9,10}. These earlier results, despite large uncertainties of up to $\pm 5 \text{ mm yr}^{-1}$, were interpreted to mean that levels of seismic hazard in the central USA should be revised downwards¹⁰.

GAMA was installed in the mid- to late 1990s and currently comprises 11 permanent geodetic monuments that both surround and straddle active faults within the New Madrid seismic zone. GAMA sites in the Mississippi embayment use a 'strong' monument consisting of a $\sim 20\text{-m}$ -long, 36-cm-diameter H-beam driven vertically into the ground with a $\sim 1\text{-m}$ mast permanently mounted on the top of the H-beam. A 'strong' monument is one where stability against small soil movements relies on the strength of the monument,

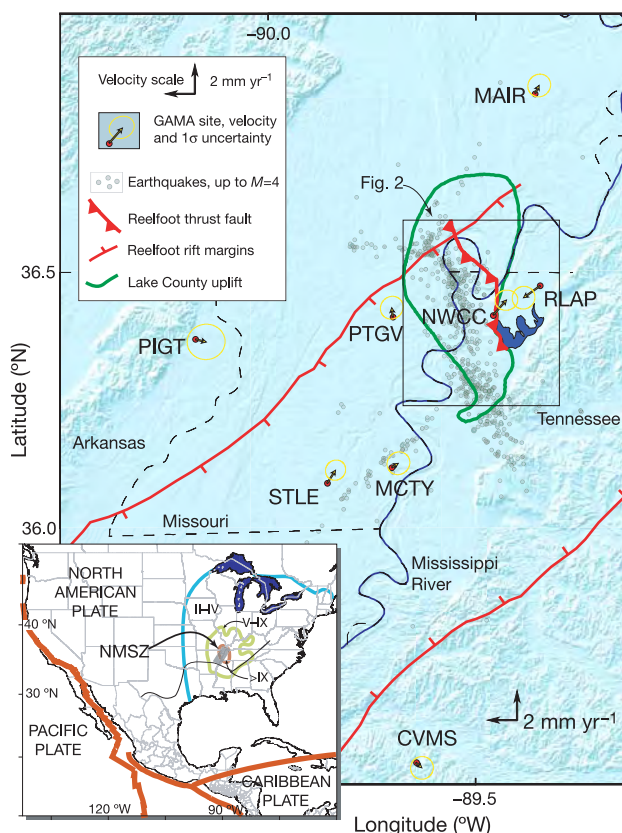


Figure 1 | Velocities and associated uncertainties of GAMA sites in the New Madrid seismic zone (NMSZ). Regional setting of the NMSZ (inset), where plate boundaries (red lines), are clearly remote. The significance of the 1811–1812 and similar earthquakes over the past 10,000 years is shown by reference to contours of intense ground-shaking, quantified by the modified Mercalli intensity scale (Roman numerals). The thick grey line under the region of highest shaking intensity is Reelfoot rift, a failed arm of the Precambrian rifted margin of North America, which is largely coincident with the interior extent of the Paleozoic Appalachian–Ouachita mountain belt (thin black line).

Table 1 | Locations, velocities, and standard uncertainties of GAMA sites

Station	Longitude	Latitude	V_E	V_N	σ_E	σ_N	Corr
RLAP	270.66	36.47	-1.13	-0.81	0.32	0.32	0.040
MAIR	270.64	36.85	0.27	0.54	0.32	0.32	0.067
NWCC	270.54	36.42	0.89	1.01	0.32	0.28	0.048
CVMS	270.36	35.54	0.40	-0.35	0.32	0.32	0.021
PTGV	270.30	36.41	-0.07	0.64	0.32	0.32	0.073
MCTY	270.30	36.12	0.44	0.28	0.32	0.32	0.070
STLE	270.14	36.09	0.57	0.89	0.28	0.28	0.069
PIGT	269.83	36.37	0.72	-0.18	0.53	0.49	0.016
GODE	283.17	39.02	-0.52	0.79	0.35	0.32	0.242
NLIB	268.43	41.77	-0.52	0.30	0.32	0.32	0.202
MDO1	255.99	30.68	0.53	0.42	0.28	0.28	0.476
PIE1	251.88	34.30	-0.35	0.54	0.32	0.32	0.561

The top eight stations are GAMA sites shown in Fig. 1; the lower four are stations used to define a stable North America. Velocities (V) and uncertainties (σ) are in mm yr^{-1} and are derived from four years of data collection. Corr is the correlation between uncertainties in the N and E directions.

¹Center for Earthquake Research and Information, ²Department of Earth Sciences, The University of Memphis, Memphis, Tennessee 38152, USA.

in this case the H-beam. This is in contrast to a braced monument, which depends on a set of braces to stabilize an otherwise 'weak' monument. Drilled braced monuments are typically constructed of ~ 2.5 -cm-diameter stainless steel rods, which are strong in compression but weak with respect to bending. Both kinds of monument typically reach depths of 10 to 15 m. (We are collaborating in a monument stability test whereby a drilled braced monument is installed within 10 m of the H-beam monument at two of the GAMA sites.) To prevent very shallow surface effects, such as frost heaving, from affecting the position of the monument the top ~ 1 m of the H-beam is decoupled from the shallow soil with a PVC pipe. A choke-ring antenna with radome is mounted on this mast. GAMA sites outside the embayment are mounted directly in rock outcrop using a ~ 3 m steel mast, the bottom ~ 2 m of which is cemented into the rock.

Velocities are derived from processing up to four years of continuous GPS data that includes the GAMA stations and additional stations in central and eastern North America (Table 1). Time-series data were processed using the GAMIT/GLOBK software package by the three-step method described by ref. 11. Formal errors are scaled by the square root of the residual chi-square per degrees of freedom to obtain the standard $1\text{-}\sigma$ uncertainty of GPS velocities¹², and a random walk of $1\text{ mm yr}^{-1/2}$ was assumed to account for possible monument instability¹³.

Two features of the distribution of surface velocities are particularly significant. First, sites close to active faults (near-field) show statistically significant motions consistent with the expected sense of displacement (Fig. 1). Two sites, NWCC and RLAP, straddle the Reelfoot thrust fault-scarp across a fault-normal distance of ~ 11 km, and show a relative convergence of $\sim 2.7 \pm 1.6\text{ mm yr}^{-1}$ (Fig. 2). The Reelfoot fault-scarp separates a region of relatively higher elevations in its hanging wall (the Lake County Uplift) from the submerged swamps of Reelfoot Lake in its footwall (Fig. 3)¹⁴. Active convergence across this fault is consistent with independent evidence for deformation associated with the fault during the third and largest of the 1811–1812 New Madrid earthquakes and for earlier Holocene activity^{15–17}. (It was displacement across this fault that notoriously caused part of the Mississippi River to flow temporarily backwards².)

Two other sites, STLE and MCTY, face each other across the

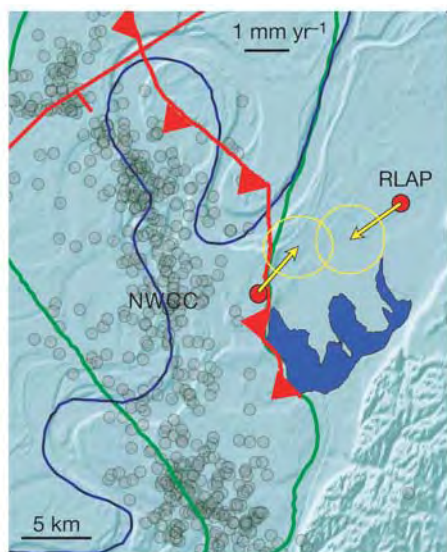


Figure 2 | Velocities of two GAMA sites, RLAP and NWCC, that straddle the active Reelfoot thrust fault. Standard 1-sigma uncertainties (see text) are shown as yellow ellipses. The thrust fault dips at $\sim 30^\circ$ to the southwest and west and is shown by the red-barbed line. Other symbols as in Fig. 1.

southern right-lateral fault, highlighted by a prominent northeast-trending and vertical zone of microseismicity and right-lateral earthquake focal mechanisms (Fig. 1). These sites are separated by a fault-normal distance of ~ 7 km and show a relative fault-parallel right-lateral velocity of $\sim 1\text{ mm yr}^{-1}$. In each case, the relative velocities yield current strain rates of the order of 10^{-7} yr^{-1} . These rates are comparable to those found along plate margins, such as the San Andreas fault in California⁸.

The second significant result is that surface velocities at distances beyond a few fault dimensions (far-field) from active faults do not differ significantly from zero (Fig. 1, Table 1). If the New Madrid seismic zone accumulates strain in the same manner as do plate boundaries, we should expect to see significant surface velocities in the far-field as well as the near-field; this is the signal pattern of one rigid region moving past another.

The apparent absence of far-field velocities suggests one of two possibilities. First, the driving force for New Madrid style earthquakes is local rather than regional. This is a fundamentally different boundary condition than typically inferred from geodetic observations along plate boundaries, in which the lateral and relative motion of plates across a relatively thin zone of deformation provides a means of accumulating strain energy. A local driving force is likely to be related to the release of gravitational potential energy, increasingly recognized as a critical source of energy in the process of building mountains within the interior of continents (for example, ref. 18). Two models have been proposed to provide a local source of energy: deformation of a low-viscosity body within the lower crust¹⁹ or the incremental sinking of a rift-pillow²⁰, each possibly triggered by the last deglaciation²¹. In each case, however, deeper motion would be expected to yield a radial surface displacement field, for which we see no current evidence.

It is also possible that the observed pattern of surface velocities represent a long-term postseismic process following the 1811–1812 earthquakes. This explanation is consistent with patterns of post-seismic deformation following, for example, the 1999 moment magnitude $M_w = 7.1$ Hector Mine²² and the 2002 $M_w = 7.9$ Denali²³ earthquakes; in each case, near-field surface velocities are significantly larger than those in the far-field. Interpretations of these patterns differ, and include any or a combination of poroelasticity decay, viscous relaxation, or afterslip across the main rupture plane. Current theoretical models are unable to distinguish among these possibilities, largely because of significant uncertainties in earth model parameters (for example, rheology, layer thicknesses,

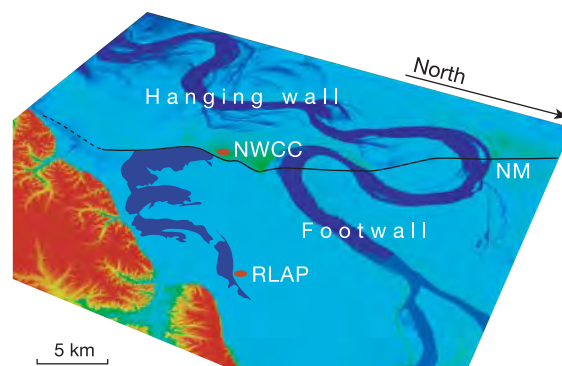


Figure 3 | An oblique view of high-resolution (10 m) digital topography associated with the Reelfoot thrust-fault. View is to the southwest and shows the relative position of the converging GAMA sites seen in Fig. 2. Surface expression of the thrust fault is shown by the black line, dashed where uncertain. The Mississippi River cuts through the clearly visible emerging hanging-wall of the Reelfoot thrust fault, and the town of New Madrid (NM) lies immediately in the footwall of the fault. The Reelfoot fault hanging wall is nowhere more than 10 m above the surrounding region and slopes gently towards the southwest.

boundary conditions) and because observational data are generally too sparse²⁴. For this reason, we have intentionally chosen not to model these data, taking the view that at this stage, modelling is premature, offering a deceptively simple and attractive solution to a complex problem. What we can say with some certainty, however, is that whatever the driving force behind the current surface velocities, whether related to 1811–1812 postseismic processes or to the accumulation of a locally sourced strain, aseismic slip is almost certainly required across faults (or shear zones) within the upper few kilometres of the surface.

A process of postseismic afterslip associated with the 1811–1812 New Madrid earthquakes is appealing, despite the relatively long time span since the events. If coseismic slip was largely confined to the subsurface, as in the analogous $M_w = 7.7$ Bhuj earthquake in Gujarat, India, in 2001, slip may be propagating into the upper few kilometres of the crust and, perhaps significantly, into relatively unconsolidated and weakly confined embayment sediments.

The new results presented here should significantly inform the discussion on the nature of deformation in the New Madrid region. Despite the large uncertainties of the earlier campaign surveys^{9,10}, those results were taken to indicate a significantly reduced level of seismic hazard in the New Madrid region. This interpretation was strongly debated^{3,6,25–29}, largely because of the extensive and unequivocal evidence for repeated large earthquakes over the past 2,000 years. Geological evidence now exists for widespread and intense liquefaction, similar in size to that generated by the 1811–1812 sequence, in AD 1450 $\pm$ 100 yr, AD 900 $\pm$ 100 yr, AD 300 $\pm$ 200 yr, and in 2350 BC $\pm$ 200 yr, and for each event, earthquakes induced more than one episode of liquefaction^{4–6,30}. We emphasize here that regardless of the geodetic results, the challenge remains to reconcile the geodetic observations with the detailed geological evidence available for repeated large earthquakes within the central USA. How such earthquakes happen inside a plate interior is not understood.

Received 4 February; accepted 11 April 2005.

- Nuttli, O. W. The Mississippi Valley earthquakes of 1811–1812: intensities, ground motion and magnitudes. *Bull. Seismol. Sci. Am.* **63**, 227–248 (1973).
- Johnston, A. C. & Schweig, E. S. The enigma of the New Madrid earthquakes of 1811–1812. *Annu. Rev. Earth Planet. Sci.* **24**, 339–384 (1996).
- The 2000 New Madrid Source workshop: reassessing New Madrid. *Eos* **81**, 397–403 (2000).
- Tuttle, M. P. & Schweig, E. S. Archeological and pedological evidence for large prehistoric earthquakes in the New Madrid seismic zone, central United States. *Geology* **23**, 253–256 (1995).
- Tuttle, M. P. The use of liquefaction features in paleoseismology: Lessons learned in the New Madrid seismic zone, central United States. *J. Seismol.* **5**, 361–380 (2001).
- Tuttle, M. P. et al. The earthquake potential of the New Madrid seismic zone. *Bull. Seismol. Soc. Am.* **92**, 2080–2089 (2002).
- Johnston, A. C. & Nava, S. J. Recurrence rates and probability estimates for the New Madrid Seismic Zone. *J. Geophys. Res.* **90**, 6737–6753 (1985).
- Hudnut, K. W., Bock, Y., Galetzka, J. E., Webb, F. H. & Young, W. H. in *Seismotectonics in Convergent Plate Boundaries* (eds Fujinawa, Y. & Yoshida, A.) 167–189 (Terrapub, Tokyo, 2002).
- Weber, J. et al. Estimation of intraplate strain accumulation in the New Madrid seismic zone from repeat GPS surveys. *Tectonics* **17**, 250–266 (1998).
- Newman, A. et al. Slow deformation and lower seismic hazard at the New Madrid Seismic Zone. *Science* **284**, 619–621 (1999).
- McClusky, S. et al. Global positioning system constraints on plate kinematics and dynamics in the eastern Mediterranean and Caucasus. *J. Geophys. Res.* **105**, 5695–5719 (2000).
- Battaglia, M., Murray, M. H., Serpelloni, E. & Bürgmann, R. The Adriatic region: an independent microplate within the Africa-Eurasia collision zone. *Geophys. Res. Lett.* **31**, doi: 10.1029/2004GL019723 (2004).
- Langbein, J. & Johnson, H. Correlated errors in geodetic time series: Implications for time-dependent deformation. *J. Geophys. Res.* **102**, 591–604 (1997).
- Russ, D. P. Style and significance of surface deformation in the vicinity of New Madrid, Missouri. *US Geol. Surv. Prof. Pap.*, 1236-H (1982).
- Kelson, K. I., Simpson, G. D., Van Arsdale, R. B., Haraden, C. C. & Lettis, W. R. Multiple late Holocene earthquakes along the Reelfoot Fault, central New Madrid seismic zone. *J. Geophys. Res.* **101**, 6151–6170 (1996).
- Fuller, M. L. *The New Madrid Earthquake* 1–119 (US Geol. Surv. Bull. 494, US Geological Survey, Washington/Denver, 1912).
- Van Arsdale, R. B., Stahle, D. W., Cleaveland, M. K. & Guccione, M. J. Earthquake signals in tree-ring data from the New Madrid Seismic Zone and implications for paleoseismicity. *Geology* **26**, 515–518 (1998).
- England, P. & Molnar, P. Active deformation of Asia; from kinematics to dynamics. *Science* **278**, 647–650 (1997).
- Kenner, S. J. & Segall, P. A mechanical model for intraplate earthquakes; application to the New Madrid seismic zone. *Science* **289**, 2329–2332 (2000).
- Pollitz, F. F., Kellogg, L. & Burgmann, R. Sinking mafic body in a reactivated lower crust: A mechanism for stress concentration at the New Madrid Seismic Zone. *Bull. Seismol. Soc. Am.* **91**, 1882–1897 (2002).
- Grollimund, B. & Zoback, M. D. Did deglaciation trigger intraplate seismicity in the New Madrid seismic zone? *Geology* **29**, 175–178 (2001).
- Hudnut, K. W. et al. Continuous GPS observations of postseismic deformation following the 16 October 1999 Hector Mine, California, earthquake (M_w 7.1). *Bull. Seismol. Soc. Am.* **92**, 1403–1422 (2002).
- Freed, A. et al. Deep lithospheric mantle and heterogeneous crustal flow following the 2002 Denali, Alaska earthquake. *Eos* **85** (Fall Meet. Suppl.), abstr. G12A–03 (2004).
- Segall, P. Postseismic Deformation: Different mechanisms in different times and places. *Eos* **85** (Fall Meet. Suppl.), abstr. G12A–01 (2004).
- Zoback, M. D. Seismic hazard at the New Madrid seismic zone; discussion and reply. *Science* **285**, 664 (1999).
- Schweig, E. S., Gombert, J. S. & Tuttle, M. P. Forum comment: Caution urged in revising earthquake hazard estimates at the New Madrid Seismic Zone. *Eos* **80**, 197 (1999).
- Newman, A. et al. Forum reply: New results justify open discussion of alternative models. *Eos* **80**, 197, 199 (1999).
- Stein, S. et al. Should Memphis build for California's earthquakes? *Eos* **84**(177), 184–185 (2003).
- Stein, S. & Newman, A. Characteristic and uncharacteristic earthquakes as possible artifacts: applications to the New Madrid and Wabash seismic zones. *Seismol. Res. Lett.* **75**, 173–187 (2004).
- Tuttle, M. P. et al. Evidence for New Madrid earthquakes in A. D. 300 and 2350 B. C. at the Burkett archeological site. *Seismol. Res. Lett.* (in the press).

Acknowledgements We thank the NSF Mid-America Earthquake Center and the US Geological Survey for supporting this work. We thank colleagues, particularly A. Johnston, at CERI and E. Schweig at the USGS for discussions.

Author Contributions R.S. and M.A.E. jointly wrote the paper and designed, constructed, and maintained the GAMA network. J.P. performed the GPS data analysis, and R.B.V.A. contributed to the interpretations.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.A.E. (mellis@memphis.edu).

First evidence of a venom delivery apparatus in extinct mammals

Richard C. Fox¹ & Craig S. Scott¹

Numerous non-mammalian vertebrates have evolved lethal venoms to aid either in securing prey or as protection from predators, but modern mammals that use venoms in these ways are rare, including only the duck-billed platypus (*Ornithorhynchus*), the Caribbean *Solenodon*, and a few shrews (Soricidae) (Order Insectivora)¹. Here we report evidence of a venom delivery apparatus in extinct mammals, documented by well-preserved specimens recovered from late Palaeocene rocks in Alberta, Canada^{2,3}. Although classified within Eutheria, these mammals are phylogenetically remote from modern Insectivora⁴ and have evolved specialized teeth as salivary venom delivery systems (VDSs) that differ markedly from one another and from those of *Solenodon* and shrews. Our discoveries therefore show that mammals have been much more flexible in the evolution of VDSs than previously believed, contradicting currently held notions that modern insectivorans are representative of the supposedly limited role of salivary venoms in mammalian history. Evidently, small predatory eutherians have paralleled colubroid snakes⁵ in evolving salivary venoms and their delivery systems several times independently.

Evidence of salivary venoms in ancient mammals is documented here in specimens of the pantolestid *Bisonalveus browni*⁶, a small North American eutherian classified in the order Cimolesta and possibly distantly related to pangolins and carnivorans⁴. The specimens were recovered from DW-2 and Birchwood, localities from the late Palaeocene Epoch (about 60 million years old⁷) in the Paskapoo Formation of central Alberta that have already yielded rich assemblages of exceptionally well preserved mammalian fossils^{2,3}. Before this study, *B. browni* was known only from poorly preserved, incomplete specimens from other parts of the western interior of North America^{6,8}; however, the best-preserved elements of *B. browni* from Alberta contain not only diagnostic posterior premolars and molars but also the incisors, canines and anterior premolars, none of which were known previously in *B. browni*. The discovery of the anterior dentition of *B. browni* not only greatly increases the knowledge of the anatomy of this peculiar mammal but also reveals an important and unexpected part of its palaeobiology—namely, a highly adapted venom delivery system, the first reported for an extinct mammal of any age.

The most informative specimen of *B. browni*, UALVP 43114 from DW-2, is an incomplete skull and lower jaws (Fig. 1). The molars on UALVP 43114 closely resemble those of *B. browni* from other localities in the western interior^{6,8}, leaving little doubt of its taxonomic identity: the teeth are massive and bunodont, the paracones are enlarged and labial in position, molar lingual cingula are well developed and the fourth premolars are molariform. Unlike previously discovered specimens of *B. browni*, the canines are preserved in place in the upper and lower jaws of UALVP 43114, with the upper canine modified to form a venom-conducting tooth: it has a robust, elongate, dagger-like crown that is sharply pointed and virtually

straight in lateral view but curves slightly lingually. Its posterior side narrows to a low carina that terminates in a low basal heel. Its anterior side is much broader and is subdivided by a wide, deep groove that extends from the base of the crown to its tip (Figs 1 and 2a–c). The walls of the groove are enamel-covered, showing that this feature is not a product of post-mortem subaerial exposure and splitting of the enamel, nor is it otherwise a taphonomic artefact of preservation. The groove is V-shaped in cross-section near the apex, but towards the root it expands into ampoule-like, C-shaped contours, while remaining open along its length (Fig. 2a–c). Moreover, this morphology is matched in all details by four isolated canines collected from DW-2 and two from Birchwood that are clearly conspecific with UALVP 43114 (Fig. 2d), indicating that the groove is not a pathological or teratological deformity affecting only one individual, but is instead an integral part of upper canine morphology in *B. browni*. From the mechanical constraints imposed by its dimensions, the groove in these teeth would have acted as a gutter, conducting fluid from its source in glandular tissues in the upper jaw down the height of the crown to its tip. No similar canine specializations have previously been described in any mammal, living or extinct, although

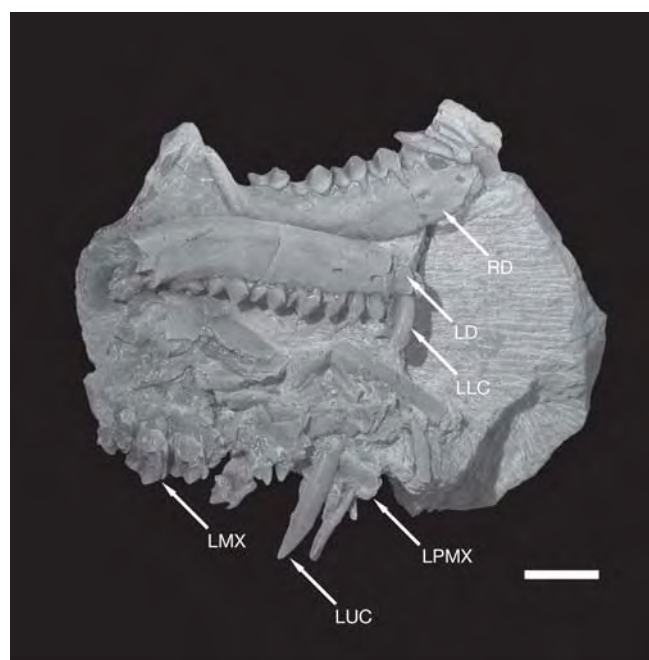


Figure 1 | *Bisonalveus browni*, specimen UALVP 43114. Incomplete skull and lower jaws (anterior to right); left upper dentition shown in lingual view. LD, left dentary; LLC, lower left canine; LMX, left maxilla; LPMX, left premaxilla; LUC, left upper canine; RD, right dentary. Scale bar, 5 mm.

¹Laboratory for Vertebrate Paleontology, Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada.

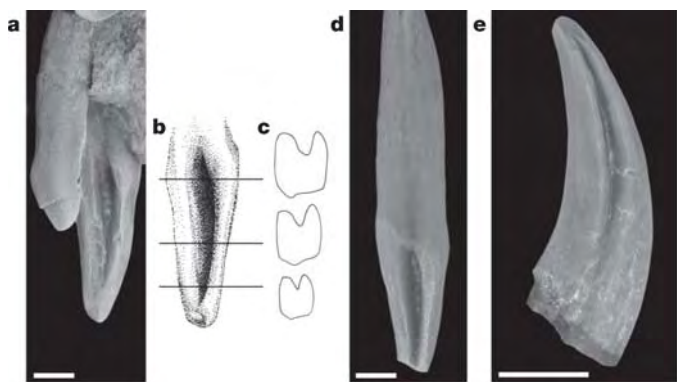


Figure 2 | *Bisonalveus browni*. **a**, UALVP 43114, left upper third incisor (at left) and canine (at right) in anterior views showing venom delivery groove in the canine. **b**, Drawing of upper canine in **a**. **c**, Sections through upper canine showing cross-dimensions of venom delivery groove along its length. **d**, UALVP 43115, isolated left upper canine from locality DW-2 in anterior view, showing venom delivery groove. Scale bars in **a** and **d**, 1 mm. **e**, Mammalia, indeterminate. Isolated right lower canine, UALVP 43116 from locality Cochrane 2, showing labial venom delivery groove; the root of this tooth is missing. Scale bar, 5 mm.

in opisthoglyphic colubroid snakes such as *Dispholidus typus*, the boomslang, the poison fangs, which are housed in the upper jaw, are also grooved anteriorly and function as a specialized VDS^{5,9,10}.

In UALVP 43114 the lower canine is highly modified, but in ways very different from the upper. Primitively in mammals, the lower canine, like the upper, is subconical, elongate, sharply pointed and somewhat recurved: when the mandible is adducted and the molars are brought into occlusion, the posterior side of the lower canine occludes with the anterior side of the upper. However, the lower canine in UALVP 43114 is substantially shorter and stouter than the upper, its crown is transversely expanded, and its overall shape is more premolar-like than caniniform (Fig. 1). Nonetheless, primitive occlusal relations with the upper canine are maintained: the posterior surface displays a broadly ovate wear facet, deeper ventrally than dorsally, developed from abrasion against the anterior side of the upper canine during chewing. The surface of the wear facet is smooth, with no part of the lower canine fitting into the groove on the upper when the two teeth meet in occlusion. Moreover, parallel scratches within the wear facet indicate that during chewing, the lower canine was moved obliquely dorsomedially across the upper and hence obliquely relative to the long axis of the groove. Given this trajectory, the groove in the upper canine did not have a mechanical function in mastication during canine occlusion in *B. browni*.

Among living mammals, only shrews and *Solenodon* use salivary venoms for immobilizing their prey, which consists of small invertebrates and even, occasionally, vertebrates¹. In the most-studied venomous species, the North American short-tailed shrew *Blarina brevicauda*¹¹, the saliva is highly toxic but none of the teeth possess clearly specialized structures to transmit the saliva to the flesh of the prey. Transmission evidently takes place during a series of rapid bites, with whatever teeth that puncture the cuticle or skin having the potential to introduce the toxic saliva into body tissues. In *Solenodon*, however, the lower second incisor is enlarged and its crown is deeply grooved posterolingually^{12,13}, thereby serving as a specialized structure for the concentrated delivery of venom, and providing a close morphological analogue in a living mammal to the grooved upper canine in *B. browni*. Like the groove in *B. browni*, the incisor groove in *Solenodon* is lined with enamel, is open along its entire length, and widens towards the base of the crown, although more so than in the canine of *B. browni*. In *Solenodon*, an enlarged submandibular salivary gland at the base of the incisor releases the venom into the incisor groove^{1,12,13}, but there is no fossa developed on the mandible

at the base of the tooth to house the gland¹⁴, nor is a fossa evident near the base of the upper canine in *B. browni*. Perhaps a specialized parotid salivary gland, nearest to the upper jaw in mammals, was the source of the venomous saliva in *B. browni*.

In *B. browni*, the difference in length and shape of the upper and lower canines seem to have important functional implications with respect to the VDS. The greater length of the upper canine would have allowed this tooth to act as a venom-conducting, piercing fang, analogous to the enlarged fangs in venomous snakes. The shorter lower canine would have enabled the tip of the upper to be more readily cleared from the lower when the mandible was depressed, thereby enhancing use of the upper canine in stabbing prey. Occlusal wear on the anterior side of the upper canine in UALVP 43114 and the isolated specimens is greatest near the base of the crown; hence, when the postcanine teeth neared full occlusion, much of the more ventral parts of the upper canine remained free and the venom groove remained open. It seems that at no time could the lower canine have obstructed the groove along its entire length and blocked the flow of venom. The wear on the opposing surfaces of the canines also suggests the potential for infusion of venom into the tissues of the prey during chewing, analogous to the chewing action that enables the envenoming of prey by many rear-fanged snakes and the Mexican/American beaded lizards, *Heloderma*¹⁵. Considering the dentition of *B. browni* in its entirety, we find it consistent with that of a small predator, with the occlusal surfaces of the posterior premolars and molars providing suitable platforms, basins, cusps and crests for the breaking up and crushing of the bodies of small arthropods and other invertebrates that had been immobilized by salivary venom delivered by the uniquely grooved upper canines.

The restriction of salivary venoms to only a few insectivorans among living mammals has posed an enigma for evolutionary biologists: why is this adaptive strategy so rare in mammals in comparison with other vertebrates, and when did it evolve? This problem has been further compounded by what has been, until this study, a complete lack of fossil evidence indicative of VDSs having a deeper history in Mammalia. As shown by our study on *B. browni*, however, mammals have been more labile in evolving specialized systems for the delivery of salivary venoms than is evident from living insectivorans alone. Moreover, our continuing research indicates that early venomous mammals might have been still more widespread than this: several isolated but well-preserved mammalian canine teeth (Fig. 2e) recovered from another late Palaeocene locality, Cochrane 2 in southern Alberta², also possess deep gutter-like grooves. These teeth are significantly larger than the upper canine of *B. browni* and are sharply pointed, recurved, bilaterally compressed and single-rooted. The groove, which has steeper walls and is deeper than that in *B. browni*, extends along the entire labial side of the crown. As indicated by a well-developed posterolabial wear surface, these teeth belong to the lower dentition, although the species to which they pertain is unknown. Plainly, the VDS of this species is not homologous to that of *B. browni*, *Solenodon* or shrews, indicating independent origins of salivary venoms in separate mammalian clades; if so, multiple origins of salivary venoms and VDSs as seen in colubroid snakes⁵ are also part of mammalian history. Specialized VDSs in early mammals may be still more anatomically and taxonomically diverse than reported here but are as yet unknown because of the generally poor preservation and scarcity in collections of the anterior dentitions of small Mesozoic and early Tertiary mammals.

Received 14 February; accepted 15 April 2005.

1. Dufton, M. J. Venomous mammals. *Pharmac. Ther.* **53**, 199–215 (1992).
2. Fox, R. C. in *Dawn of the Age of Mammals in the Northern Part of the Rocky Mountain Interior* (eds Bown, T. M. & Rose, K. D.) 51–70 (Geological Society of America, Boulder, Colorado, 1990).
3. Scott, C. S. Taxonomically diverse late Paleocene mammal localities from south central Alberta, Canada. *J. Vertebr. Paleont.* **24**, 111A (2004).

4. McKenna, M. C. & Bell, S. K. *Classification of Mammals Above the Species Level* (Columbia Univ. Press, New York, 1997).
 5. Jackson, K. The evolution of venom-delivery systems in snakes. *Zool. J. Linn. Soc.* **137**, 337–354 (2003).
 6. Gazin, C. L. Paleocene mammalian faunas of the Bison Basin in south-central Wyoming. *Smithson. Misc. Coll.* **131**, 1–57 (1956).
 7. Lofgren, D. L., et al. in *Late Cretaceous and Cenozoic Mammals of North America* (ed. Woodburne, M. O.) 43–105 (Columbia Univ. Press, New York, 2004).
 8. Gingerich, P. D. New Adapisoricidae, Pentacodontidae, and Hyopsodontidae (Mammalia, Insectivora and Condylarthra) from the late Paleocene of Wyoming and Colorado. *Contrib. Mus. Paleont. Univ. Mich.* **26**, 227–255 (1983).
 9. Bellairs, A. d'A. *Reptiles: Life History, Evolution, and Structure* (Harper, New York, 1957).
 10. Lake, A. R. & Trevor-Jones, T. R. The venom apparatus of the boomslang or tree snake, *Dispholidus typus*. *S. Afr. J. Sci.* **92**, 167–169 (1996).
 11. Kita, M. et al. Blarina toxin, a mammalian lethal venom from the short-tailed shrew, *Blarina brevicauda*: isolation and characterization. *Proc. Natl Acad. Sci. USA* **101**, 7542–7547 (2004).
 12. Rabb, G. B. Toxic salivary glands in the primitive insectivore *Solenodon*. *Nat. Hist. Miscellanea* **190**, 1–3 (1959).
 13. Macdonald, D. (ed.) *The Encyclopedia of Mammals* (Facts On File, New York, 1984).
 14. McDowell, S. B. Jr The Greater Antillean insectivores. *Bull. Am. Mus. Nat. Hist.* **115**, 113–214 (1958).
 15. Goin, C. J. & Goin, O. B. *Introduction to Herpetology* (W. H. Freeman, San Francisco, 1972).
- Acknowledgements** We thank Y. Sun for fossil preparation, D. Spivak for technical assistance, M. Caldwell for use of photographic equipment, and K. Gao, K. Soehn, G. Stonley, M. Webb and G. Youzwysyn for field assistance. This work was supported by the Natural Sciences and Engineering Research Council of Canada and the University of Alberta.
- Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.C.F. (richard.fox@ualberta.ca).

LETTERS

Aerodynamics of the hovering hummingbird

Douglas R. Warrick¹, Bret W. Tobalske² & Donald R. Powers³

Despite profound musculoskeletal differences, hummingbirds (Trochilidae) are widely thought to employ aerodynamic mechanisms similar to those used by insects. The kinematic symmetry of the hummingbird upstroke and downstroke^{1–3} has led to the assumption that these halves of the wingbeat cycle contribute equally to weight support during hovering, as exhibited by insects of similar size⁴. This assumption has been applied, either explicitly or implicitly, in widely used aerodynamic models^{1,5–7} and in a variety of empirical tests^{8,9}. Here we provide measurements of the wake of hovering rufous hummingbirds (*Selasphorus rufus*) obtained with digital particle image velocimetry that show force asymmetry: hummingbirds produce 75% of their weight support during the downstroke and only 25% during the upstroke. Some of this asymmetry is probably due to inversion of their cambered wings during upstroke. The wake of hummingbird wings also reveals evidence of leading-edge vortices created during the downstroke, indicating that they may operate at Reynolds numbers sufficiently low to exploit a key mechanism typical of insect hovering^{10,11}. Hummingbird hovering approaches that of insects, yet remains distinct because of effects resulting from an inherently dissimilar—avian—body plan.

The convergence of hummingbirds and nectivorous insects in the form and use of wings is a testament to the strength of the selective forces imposed: the demands of a high-energy flux way of life, and of locomotion in a fluid medium. The strong similarities in morphology and kinematics—and hence in Reynolds (Re) numbers—have led to the prediction that, during hovering flight in particular, hummingbirds and insects produce lift by using similar aerodynamic mechanisms^{1,3,12,13}. Studies of the flight of hawkmoths (Re ≈ 8000) suggest twofold similarities: the use of symmetrical hovering, where the upstroke and downstroke contribute roughly equally to weight support; and the use of aerodynamic mechanisms previously unrecognized in birds, including dynamic stall and leading-edge vortices, to augment lift^{10,11,14–16}. The purpose of the present study was to test, with the use of digital particle imaging velocimetry (DPIV), whether hummingbirds have completely converged with insects in terms of the aerodynamic contributions of the upstroke and downstroke, and to examine the wake structure for evidence that hummingbirds exploit low-Re aerodynamic mechanisms characteristic of insect flight.

We sampled the wake produced by rufous hummingbirds (*Selasphorus rufus*, $n = 3$; Re during hover ≈ 3,000, average wing chord as characteristic length) as they hovered in the 60 × 60 × 85 cm³ test section of an idle wind tunnel with the bottom removed. The wake structure was measured less than four wing chord lengths (4 cm) below the wings, 8–14 ms after the kinematic events that produced it, and before the wake structure significantly degraded owing to viscosity or near-field interaction with the still-active wings¹⁷. The birds were trained to fly to a feeder (a 1-ml syringe containing a 20% sucrose solution), the placement of which was manipulated to allow two-dimensional sampling of different

portions of the wake from up to three chord lengths (3 cm) above the bird to eight chord lengths (8 cm) below the bird. Simultaneous digital video (500 frames s^{−1}), taken from directly above the bird, allowed the synchronization of DPIV images with wing kinematics. The wake was sampled in five frontal planes (Fig. 1a, b) and parasagittal planes (Fig. 1c, d) taken at 1-cm intervals.

Frontal-plane samples of the wake revealed trailing-tip vortices as the wings passed in and out of the plane of the laser sheet (Fig. 1b); parasagittal planes revealed the structure of the starting and stopping vortices of the downstroke, the starting vortices of the upstroke, and any leading-edge vorticity (LEV) shed at the end of the halfstroke¹⁰ (Fig. 1d). For both sampling planes, only images taken during the transition from upstroke to downstroke revealed distinct structure created by the immediately previous upstroke and downstroke, respectively. At all other periods in the wingbeat cycle, the dominant momentum jet produced by the downstroke annihilated any major vortex structures produced by the upstroke. Sixty-five images were

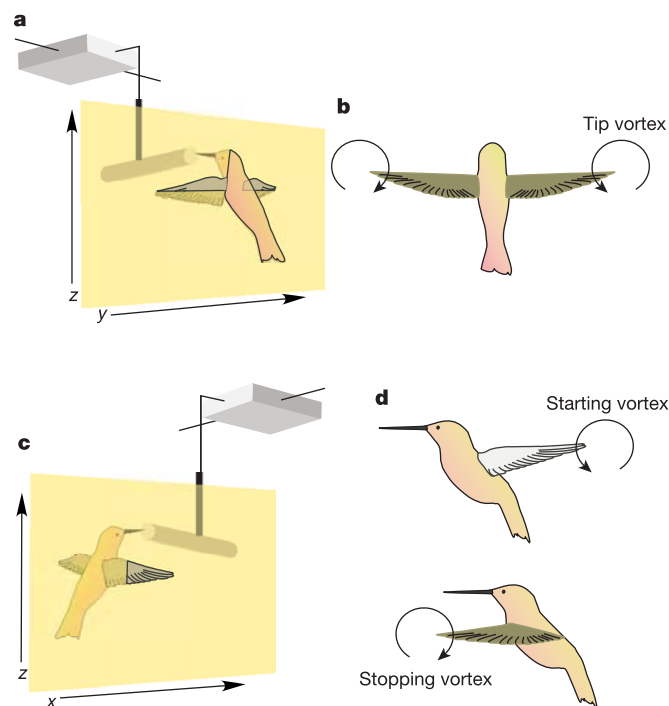


Figure 1 | DPIV methodology. **a**, Birds hovering at a feeder positioned such that the two-dimensional laser light sheet illuminated wake-entrained oil particles moving in a frontal plane. **b**, The expected wake structures within the frontal plane^{5,6}. **c**, Light sheet oriented in the parasagittal plane, centred at midwing. **d**, The expected wake structures within the parasagittal plane^{5,6}.

¹Department of Zoology, Oregon State University, 3029 Cordley Hall, Corvallis, Oregon 97331, USA. ²Department of Biology, University of Portland, 5000 North Willamette Boulevard, Portland, Oregon 97203, USA. ³Biology Department, George Fox University, 414 North Meridian Street, Newberg, Oregon 97132, USA.

analysed for the aerodynamic contribution of the two halves of the wingbeat cycle, and paired comparisons of the upstroke and downstroke tip vortices (frontal planes; Fig. 2a) and downstroke stopping and upstroke starting vortices (parasagittal planes (Fig. 2b) were made.

Consistent with their steady position in hovering was the observation that there was sufficient circulation, Γ , in the tip vortices, and adequate separation between vortex cores, to support the body weight of the hummingbirds. The circulation (mean $\pm$ s.d.) relative to that required for weight support was $111 \pm 20\%$ for bird 1 ($n = 11$), $94 \pm 17\%$ for bird 2 ($n = 25$) and $103 \pm 20\%$ for bird 3 ($n = 29$); the range for all samples was 57–147% ($n = 65$).

Eight of 15 parasagittal samples revealed wake structures consistent with the production of leading-edge vortices during the downstroke. In these samples, at the transition from downstroke to upstroke, two cores of vorticity, both rotating in the same direction, were shed in quick succession. The first (LEV_D ; Fig. 2b) is consistent with the shedding of a leading-edge vortex generated by the previous downstroke. As demonstrated by robotic models operating at similar Re (ref. 10), this vortex can remain attached to the wing beneath the bound circulation and is shed during the supinating rotation at the beginning of upstroke. Soon afterwards, the starting vortex of the upstroke is produced (U, Fig. 2b).

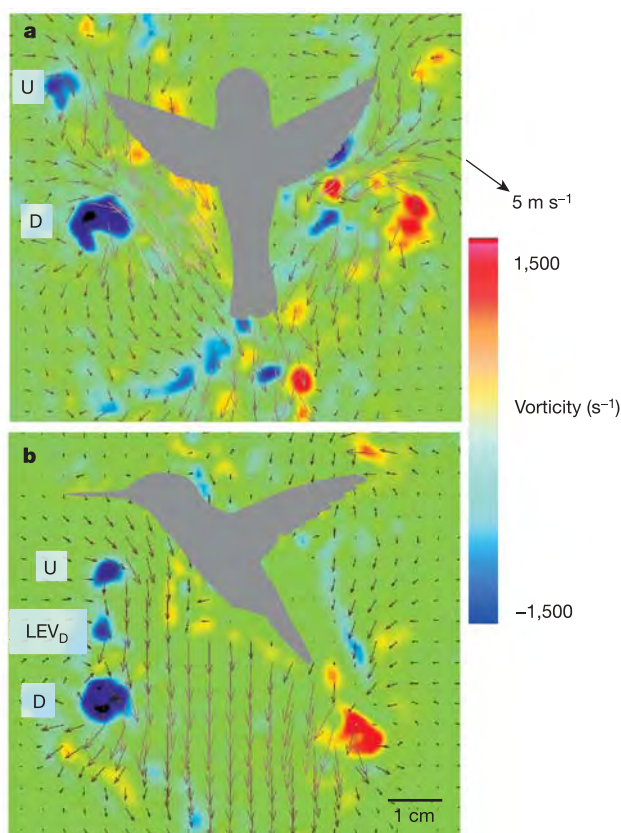


Figure 2 | Flow field vorticity. **a**, Single-field sample of a hummingbird wake taken at the end of the upstroke, in frontal view with the interrogation plane through the shoulder, revealing the tip vortices of downstroke D and upstroke U. **b**, Single-field sample of a hummingbird wake at the end of the upstroke, in parasagittal plane with the interrogation plane passing through the midwing during the middle of each half wingbeat (about 3 cm from the midline of the bird's body). Between the downstroke stopping vortex D and the upstroke starting vortex U is a pocket of vorticity LEV_D presumably created at the leading edge of the wing during the rapid wing pronation at the beginning of the preceding downstroke, and carried through the downstroke to be shed during the supination at the beginning of the upstroke.

Given the strength of this presumptive leading-edge vortex, and if we assume that it was feeding into the tip vortex strength¹⁰ (as measured in the frontal samples), the leading-edge vortex provided $15 \pm 4\%$ of the circulation of the tip vortex wake during downstroke. Because leading-edge vortices can be generated during the rapid ventral rotation (pronation) of the wing—before the downstroke, wing movement can generate a more traditional bound vortex—it is possible that a more important function of the formation of the leading-edge vortex is to draw air downwards during wing turn-around¹³, filling the gap in aerodynamic force production at wing

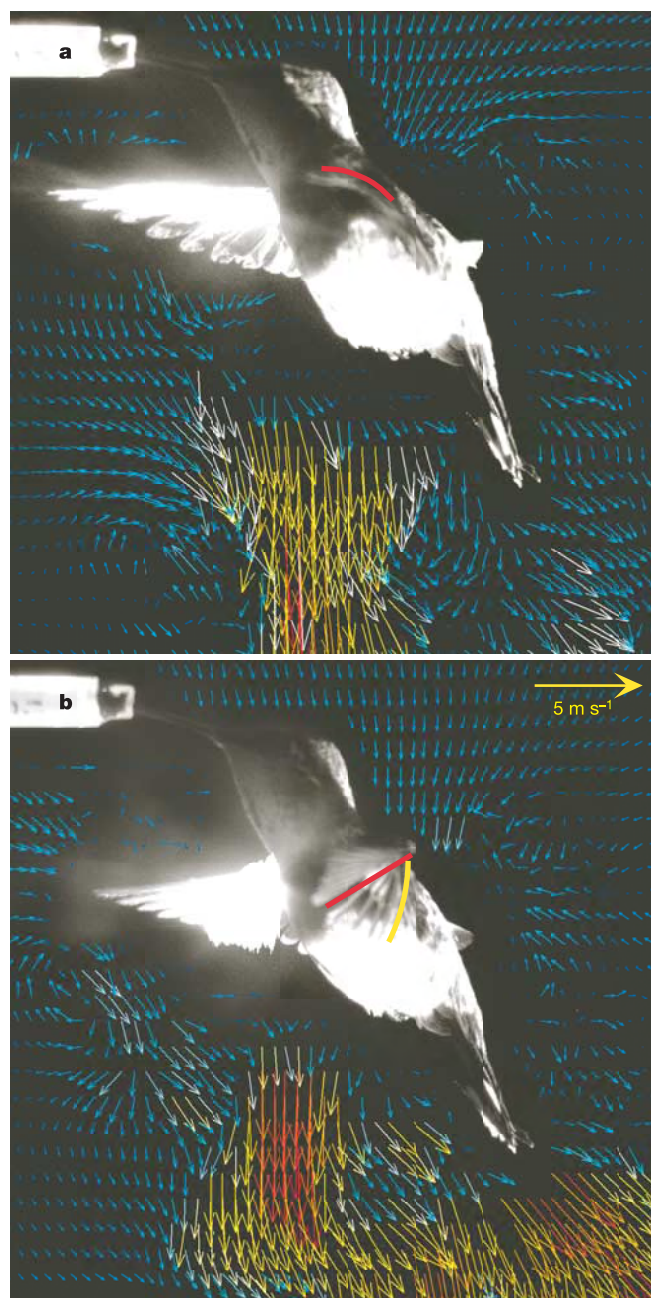


Figure 3 | Hummingbird wing presentation and flow field. **a**, The profile of the hummingbird wing during mid-downstroke. A red line is drawn just above the dorsal surface of the wing to highlight the camber, typical of a bird wing. **b**, The proximal part of the wing, to the left of the yellow line, is not as supinated (inverted) as the distal portion, to the left of the red line. The presentation of these two portions of the wing is typical for an avian upstroke, and results in an airfoil of reduced efficacy relative to the downstroke. Vector scale top right.

turnaround, and perhaps facilitating the rapid development of more typical bound circulation.

Pairwise comparisons of upstroke and downstroke circulation within a wingbeat cycle revealed that the downstroke provides the vast majority of the weight support during hovering (averages for individuals, $74 \pm 7\%$, $77 \pm 5\%$ and $76 \pm 6\%$; paired *t*-test for differences in circulation between upstroke/downstroke vortex pairs, $P < 0.0001$). Downstroke circulation varied less than that of the upstroke (coefficient of variation for all birds, all samples: 0.21 versus 0.27).

Long-standing kinematic data have suggested that hummingbird upstroke and downstroke are essentially symmetrical^{1–3}. Our own three-dimensional kinematics indicate that wing span is indeed fairly symmetrical (upstroke 91 ± 4 mm; downstroke 93 ± 3 mm). The angular velocity of the wing is slightly less during the upstroke (146 ± 15 rad s⁻¹) than during the downstroke (196 ± 26 rad s⁻¹). Observed differences in area (*S*) and velocity (*v*) applied to a general lift (*L*) formula ($L \propto v^2 S C_L$) suggest that the downstroke should produce 64% of weight support.

The remaining variable, lift coefficient (C_L), is a function of angle of attack. On average, the angle of attack (measured at midwing, halfway between the wrist and tip) at mid-downstroke is greater than during mid-upstroke ($36 \pm 12^\circ$ versus $26 \pm 13^\circ$), the higher angles during downstroke allowed by the positive camber typical of bird wings (Fig. 3a). Lift coefficients vary linearly with angle of attack; adding this difference in C_L would result in a downstroke that produced about 70% of weight support—somewhat less than the 75% value observed in the wake. The remainder of the disparity might be produced by more subtle asymmetries—in particular, the angle of attack and performance of the proximal wing, which does not seem to reverse camber during the upstroke (Fig. 3b). In addition, we saw no evidence of leading-edge vorticity either developing at the beginning of the upstroke or being shed at the end of it.

Although the wing kinematics of hummingbirds shows strong convergence with that of certain insects (for example the hawkmoth, *Manduca sexta*), there are fundamental musculoskeletal and planform material properties that limit their ability to produce the wingbeat cycle symmetry found in these insects. In comparison with other birds, the hummingbird shoulder joint allows greater rotation about the long axis of the humerus, and the relatively long primaries of hummingbirds allows much more of the wing planform to be inverted during the upstroke^{2,3,18}. Nonetheless, our results show that the efficacy of the hummingbird wing at negative angles of attack in the upstroke compares poorly with its performance at positive angles of attack during the downstroke. Although camber in the distal wing could conceivably be reversed, in the proximal wing thin feathers trailing behind thick musculoskeletal elements create camber that is not reversed during the upstroke (Fig. 3b). In contrast, the elastic qualities of insect wings in response to both aerodynamic¹⁹ and inertial forces²⁰ allow them to reverse their camber fully, and therefore to develop high lift coefficients during both half-strokes¹⁵.

Hummingbirds and insects have evolved for sustained hovering flight from vastly different ancestral directions, and their distinct phylogenies underlie the differences in their aerodynamic styles. In all other birds—and, presumably, hummingbird ancestors—the downstroke provides 100% of weight support during slow flight and hovering⁷. Given that many birds possess the mass-specific power (using anaerobic metabolism) to hover for short periods, the selective pressure on hummingbird ancestors was probably for increased efficiency^{1,8} (resulting in stiff wings with greatly simplified kinematics), and an upstroke muscle (the supracoracoideus) that makes the recovery stroke rapid, while contributing enough to the hovering power requirements to allow the downstroke muscle (the pectoralis) to operate within its aerobic limits. In other words, this pseudosymmetrical wingbeat cycle is good enough, and although

hummingbirds do not exhibit the elegant aerodynamic symmetry of insects, natural selection rewards ‘good enough’ as richly as it does our aesthetic ideals.

METHODS

DPIV. Our DPIV system was manufactured by LaVision Inc.; recording and analysis were accomplished with DaVis (v6.0.2). We used a dual-cavity pulsed 50-mJ Nd:YAG laser to illuminate a flow field about 3 mm thick, with planar dimensions spanning the wake of the hummingbird from three chord lengths above to eight chord lengths below the wing root. The air was seeded with submicrometre-sized particles of olive oil vapour, generated with a Laskin nozzle at a rate of 7×10^{10} particles s⁻¹. Particle illumination was recorded with a 1,376-pixel $\times$ 1,040-pixel charge-coupled device camera. To calculate velocity, a cross-correlation with adaptive multipass was employed; this method correlates within areas beginning at 64 pixels $\times$ 64 pixels and decreasing to 16 pixels $\times$ 16 pixels with 50% overlap. A correlation peak error of 0.1 pixel, and an average particle separation in the wake of 12 pixels, produced 1% error^{21,22}; combined with optical distortion and particle-fluid fidelity error²², our observed measurement error was $2.3 \pm 0.5\%$. To compute vorticity (ω , s⁻¹), we post-processed vector fields with a median filter and then computed $\text{rotz}(\text{dy}/\text{dx})$. Background ω , measured 2–3 cm outside the wake structure, was less than 2% of peak ω in the wake; a 10% mask was applied to eliminate this background noise and allow the definition of wake structures. The scaled pixel area of the remaining wake structure was then summed, yielding total circulation.

Circulation and weight support. We measured circulation (Γ , m² s⁻¹) in the trailing-tip vortices by integrating ω with respect to area (m²). We limited our analysis to views where vortex cores were normal to the sampling plane (parasagittal, centred at midwing; frontal, centred at wing root)²³. We tested whether observed Γ was sufficient to support body weight by comparing observed Γ with circulation required (Γ_o), where $\Gamma_o = WT/\rho S$, where *W* is body weight (N), *T* is time per wingbeat (s) and *S* is the projected horizontal area swept by the two wings (m²) (ref. 23).

Kinematics. Separate flight trials (*n* = 4 birds) were recorded with two synchronized high-speed digital video cameras operating at 500 Hz sampling and a shutter speed of 1/1,000 s. We merged two-dimensional coordinates from each camera into a single three-dimensional coordinate space by using the direct linear transformation coefficients derived from a 16-point calibration frame²⁴. From these data we calculated the horizontal projection of the stroke plane, the angular velocity of the wing (rad s⁻¹) and the angle of attack of the midwing (degrees) relative to the incident air flow. Incident air velocity was the sum of translational velocity of the wing and average three-dimensional air velocity computed with DPIV data from the frontal and sagittal planes, which is dominated by a mean downward velocity of 1.1 m s⁻¹.

Received 9 March; accepted 18 April 2005.

1. Weis-Fogh, T. Energetics of hovering flight in hummingbirds and in *Drosophila*. *J. Exp. Biol.* **56**, 79–104 (1972).
2. Stolpe, V. M. & Zimmer, K. Der Schwirrflyug des Kolibri im Zeitlupenfilm. *J. Ornithol.* **87**, 136–155 (1939).
3. Greenwalt, C. H. The wings of insects and birds as mechanical oscillators. *Proc. Am. Phil. Soc.* **104**, 605–611 (1960).
4. Willmott, A. P. & Ellington, C. P. The mechanics of flight in the hawkmoth *Manduca sexta*. II. Aerodynamic consequences of kinematic and morphological variation. *J. Exp. Biol.* **200**, 2723–2745 (1997).
5. Rayner, J. M. V. A vortex theory of animal flight. I. The vortex wake of a hovering animal. *J. Fluid Mech.* **91**, 697–730 (1979).
6. Ellington, C. P. The aerodynamics of hovering insect flight. V. A vortex theory. *Phil. Trans. R. Soc. Lond. B* **305**, 79–113 (1984).
7. Norberg, U. M. *Vertebrate Flight: Mechanics, Physiology, Morphology, Ecology, and Evolution* (Springer, Berlin, 1990).
8. Wells, D. Muscle performance in hovering hummingbirds. *J. Exp. Biol.* **78**, 39–57 (1993).
9. Tobalske, B. W., Altshuler, D. L. & Powers, D. R. Take-off mechanics in hummingbirds (Trochilidae). *J. Exp. Biol.* **207**, 1345–1352 (2004).
10. van den Berg, C. & Ellington, C. P. The vortex wake of a ‘hovering’ model hawkmoth. *Phil. Trans. R. Soc. Lond. B* **352**, 329–340 (1997).
11. Dickinson, M. H., Lehmann, F. & Sane, S. P. Wing rotation and the aerodynamic basis of insect flight. *Science* **284**, 1954–1960 (1999).
12. Dudley, R. *The Biomechanics of Insect Flight* (Princeton Univ. Press, Princeton, 2000).
13. Altshuler, D., Dudley, R. & Ellington, C. P. Aerodynamic forces of revolving hummingbird wings and wing models. *J. Zool. (Lond.)* **264**, 327–332 (2004).
14. Willmott, A. P. & Ellington, C. P. The mechanics of flight in the hawkmoth *Manduca sexta*. I. Kinematics of hovering and forward flight. *J. Exp. Biol.* **200**, 2705–2722 (1997).

15. Willmott, A. P., Ellington, C. P. & Thomas, A. L. R. Flow visualization and unsteady aerodynamics in the flight of the hawkmoth *Manduca sexta*. *Phil. Trans. R. Soc. Lond. B* **352**, 303–316 (1997).
 16. Usherwood, J. R. & Ellington, C. P. The aerodynamics of revolving wings. I. Model hawkmoth wings. *J. Exp. Biol.* **205**, 1547–1564 (2002).
 17. Tytell, E. D. & Ellington, C. P. How to perform measurements in a hovering animal's wake: physical modelling of the vortex wake of the hawkmoth, *Manduca sexta*. *Phil. Trans. R. Soc. Lond. B* **358**, 1559–1566 (2003).
 18. Tobalske, B. W., Hedrick, T. L. & Biewener, A. A. Wing kinematics of avian flight across speeds. *J. Avian Biol.* **34**, 177–184 (2003).
 19. Wootten, R. J. Geometry and mechanics of insect hindwing fans: a modelling approach. *Proc. R. Soc. Lond. B* **262**, 181–187 (1995).
 20. Combes, S. A. & Daniel, T. L. Into thin air: contributions of aerodynamic and inertial-elastic forces to wing bending in the hawkmoth *Manduca sexta*. *J. Exp. Biol.* **206**, 2999–3006 (2003).
 21. Spedding, G. R., Hedenstrom, A. & Rosen, M. Quantitative studies of the wakes of freely flying birds in a low-turbulence wind tunnel. *Exp. Fluids* **34**, 291–303 (2003).
 22. Raffel, M., Willert, C. & Kompenhans, J. *Particle Image Velocimetry: A Practical Guide* (Springer, Berlin, 2000).
 23. Spedding, G. R., Rosen, M. & Hedenstrom, A. A family of vortex wakes generated by a thrush in free flight in a wind tunnel of its entire natural range of flight speeds. *J. Exp. Biol.* **206**, 2313–2344 (2003).
 24. Hedrick, T. L., Tobalske, B. W. & Biewener, A. A. Estimates of gait change based on a three-dimensional analysis of flight in cockatiels (*Nymphicus hollandicus*) and ringed turtle doves (*Streptopelia risoria*). *J. Exp. Biol.* **205**, 1389–1409 (2002).
- Acknowledgements** We thank B. Klopfenstein for her help with the experiments. This work was supported by grants from the National Science Foundation and the Murdock Charitable Trust.
- Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.R.W. (warrickd@science.oregonstate.edu).

LETTERS

Extracellular electron transfer via microbial nanowires

Gemma Reguera¹, Kevin D. McCarthy^{2*}, Teena Mehta^{1*}, Julie S. Nicoll¹, Mark T. Tuominen² & Derek R. Lovley¹

Microbes that can transfer electrons to extracellular electron acceptors, such as Fe(III) oxides, are important in organic matter degradation and nutrient cycling in soils and sediments^{1,2}. Previous investigations on electron transfer to Fe(III) have focused on the role of outer-membrane *c*-type cytochromes^{1,3}. However, some Fe(III) reducers lack *c*-cytochromes⁴. *Geobacter* species, which are the predominant Fe(III) reducers in many environments¹, must directly contact Fe(III) oxides to reduce them⁵, and produce monolateral pili⁶ that were proposed^{1,2}, on the basis of the role of pili in other organisms^{7,8}, to aid in establishing contact with the Fe(III) oxides. Here we report that a pilus-deficient mutant of *Geobacter sulfurreducens* could not reduce Fe(III) oxides but could attach to them. Conducting-probe atomic force microscopy revealed that the pili were highly conductive. These results indicate that the pili of *G. sulfurreducens* might serve as biological nanowires, transferring electrons from the cell surface to the surface of Fe(III) oxides. Electron transfer through pili indicates possibilities for other unique cell-surface and cell-cell interactions, and for bioengineering of novel conductive materials.

The role of pili in Fe(III) oxide reduction was studied with *Geobacter sulfurreducens* because a genetic system⁹ and the complete genome sequence¹⁰ are available. As expected from previous studies⁶, *G. sulfurreducens* produced pili during growth on Fe(III) oxide

(Fig. 1a) but not on soluble Fe(III) (Fig. 1b), and the pili were localized to one side of the cell. The formation of pili could also be induced during growth on the alternative electron acceptor fumarate if the cells were grown at the suboptimal temperature of 25 °C (Fig. 2a), indicating that pilin production in *G. sulfurreducens* might be growth-regulated as it is in other bacteria¹¹.

The genome sequence of *G. sulfurreducens* contained two open reading frames (ORFs), GSU1496 and GSU1776, predicted to code for pilin domain proteins with the conserved amino-terminal amino acid characteristics of type IV pilins¹². Phylogenetic analyses placed the protein encoded by ORF GSU1776 among bacterial pseudopilins of type II secretion systems, and subsequent studies have confirmed the role of this gene, termed *oxpG*, in protein secretion to the outer membrane¹³. The protein encoded by ORF GSU1496 formed an independent line of descent along with pilin subunits of other members of the *Geobacteraceae* such as *Geobacter metallireducens* and *Pelobacter propionicus* (Fig. 1c). The predicted length of these *Geobacter* pilin proteins was considerably shorter than other bacterial pilins (see Supplementary Fig. S1) and was restricted to the highly conserved N-terminal domain of bacterial type IV pilins, which functions in inner membrane insertion, signal processing and pilin polymerization, and forms the central helical core of the pilus filament^{12,14}. The degree of conservation of geopilins at this region was lower than other bacterial pilins and, as a result, geopilins were

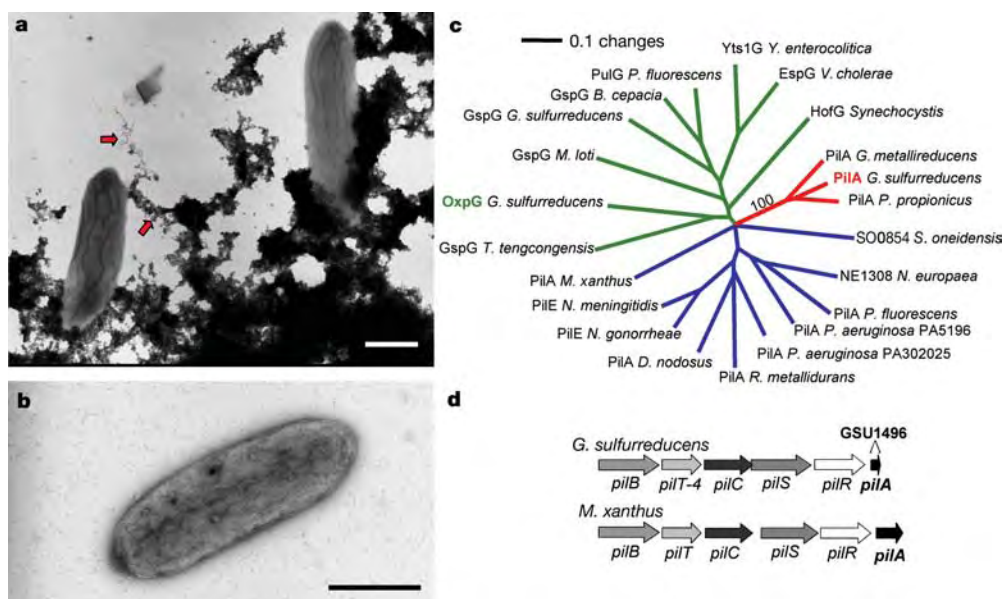


Figure 1 | *Geobacter sulfurreducens* pili. **a, b**, Transmission electron micrographs of cells grown with poorly crystalline Fe(III) oxides (**a**) or soluble Fe(III) citrate (**b**). Arrows indicate pili. Scale bars, 0.5 μ m. **c**, Phylogenetic distance tree derived from amino-terminal amino acid sequence alignments (see Supplementary Information and Supplementary Fig. S1) showing the relationship of annotated *G. sulfurreducens* pilin-domain proteins, encoded by GSU1496 and GSU1776, and pilins (blue branches) and pseudopilins (green branches) of other bacteria. **d**, Genomic organization of pilus biosynthesis genes surrounding GSU1496 in *G. sulfurreducens* and the *pilA* gene in *M. xanthus*¹⁷.

¹Department of Microbiology and ²Department of Physics, University of Massachusetts, Amherst, Massachusetts 01003, USA.

*These authors contributed equally to this work.

phylogenetically distant from other bacterial pilins, including a type IV pilin of another metal reducer, *Shewanella oneidensis* (Fig. 1c). Homologues of genes required for the formation and assembly of pili in other Gram-negative bacteria^{15,16} are upstream of the *Geobacter pilA* gene (Fig. 1d), in a genetic arrangement similar to that of the pili genes in *Myxococcus xanthus*¹⁷, a δ -proteobacterium distantly related to *Geobacter*. These results indicate that the GSU1496 gene encodes a pilin subunit, which was designated PilA.

When *pilA* was deleted, *G. sulfurreducens* failed to produce pili (Fig. 2b) and could no longer reduce insoluble electron acceptors such as poorly crystalline Fe(III) oxides (Fig. 3) and Mn(IV) oxides (data not shown). In contrast, the mutant could reduce soluble electron acceptors, such as fumarate and Fe(III) citrate as well as the wild type. The mutant also grew in medium containing Fe(III) oxide if the chelator nitrilotriacetate was added to solubilize some of the Fe(III) or in the presence of anthraquinone-2,6-disulfonate (AQDS) (Supplementary Fig. S2). AQDS serves as a soluble electron shuttle and transfers electrons between the cell surface and the surface of the Fe(III) oxide¹⁸, alleviating the need for direct contact for Fe(III) oxide reduction⁵. Complementation of the *pilA* mutation with a functional copy of the *pilA* gene *in trans* restored the capacity for assembling pili (Fig. 2c) and for Fe(III) oxide reduction (Fig. 3a). These results showed that *G. sulfurreducens* required the assembly of functional pili to reduce insoluble Fe(III) oxides.

One known function of type IV pili in other microorganisms is establishing contact with surfaces^{7,8}. Fe(III) oxides are typically smaller than *G. sulfurreducens* cells (Fig. 1a), but it was possible to quantify the potential for attachment of *G. sulfurreducens* to Fe(III) by inoculating fumarate-grown cells into medium in which Fe(III) oxide, attached to glass coverslips, was provided as the sole electron acceptor. Within the first 24 h, the cells of the *pilA*-deficient strain that were added initially attached to Fe(III) oxides as well as the wild type (Fig. 3b), but whereas the wild type grew on the Fe(III) oxide, as indicated by an increase in biomass on the Fe(III) oxide over the next 24 h, the *pilA* mutant could not grow, as shown by a decrease in biomass (Fig. 3b). The *pilA*-deficient mutant did grow on the surface if fumarate was provided as an alternative electron acceptor (data not shown). These results showed that pili are not required for Fe(III) oxides to attach to cells and confirmed the necessity for pili for growth with Fe(III) oxides as the sole electron acceptor. Further evaluation of the nature of the association of the Fe(III) oxides with the cells revealed that, when Fe(III) oxides were added to fumarate-grown cells, the outer surface of the *pilA*-deficient mutant still had the ability to bind Fe(III) oxides (Fig. 3d) but in the wild type there was substantial association of Fe(III) oxides with the pili (Fig. 3c).

It has previously been proposed that *Geobacter's* pili might mediate surface motility, which might aid *G. sulfurreducens* in locating Fe(III) or Mn(VI) oxides^{1,2,6}, but no twitching motility of the wild-type cells was observed on glass surfaces coated with Fe(III) oxide. Furthermore, deleting a putative *pilT* gene (Supplementary Information), which is required for twitching motility in other organisms¹⁹, had no effect on Fe(III) oxide reduction.

These results indicated that the pili might have a more direct role in electron transfer to Fe(III) oxides. To evaluate this we measured the electrical conductivity through the pili. Pili and other proteins released from the outer surface of *G. sulfurreducens* grown with fumarate (Supplementary Fig. S3) were immobilized on a graphite surface and analysed with an atomic force microscope (AFM) equipped with a conductive tip and electronics that permitted mapping of the local conductance from the tip to the substrate (Fig. 4). Topographic analysis revealed pili as well as other, unidentified, more globular, proteins which were also sheared off the outer cell surface (Fig. 4a). When a voltage was applied to the tip there was a strong current response along the pilus filament, which was positive when a positive voltage was applied and negative with a negative voltage (Fig. 4b, c). In contrast, the non-pilin proteins had no detectable conductivity and in instances in which the non-pilin

proteins covered the pili filaments they insulated the pili from the conductive tip. This general response, initially observed in relatively large-scale scans (Fig. 4a–c), was even clearer in cross-sections in which high current was associated with the slight increase in topography associated with the pilus filament, but the higher topography, associated with non-pilin material, had no detectable current (Fig. 4d, middle panel). A scan across a portion of the pilin filament overlain by other material also yielded no detectable current (Fig. 4d, bottom panel). Current line scans generated after applying different voltages while scanning the same region of a pilus demonstrated a linear, ohmic, correspondence between current and voltage applied (Fig. 4e, f). When similar studies were performed with pili from the metal reducer *Shewanella oneidensis* or the non-metal reducer *Pseudomonas aeruginosa* (Supplementary Fig. S4 and S5), no conductance was detected.

These results show that the pili of *G. sulfurreducens* are highly conductive. This indicates that *G. sulfurreducens* requires pili in order to reduce Fe(III) oxides because pili are the electrical connection between the cell and the surface of the Fe(III) oxides. This contrasts with the nearly universal concept that outer-membrane cytochromes are the proteins that transfer electrons to Fe(III) oxide in Fe(III)

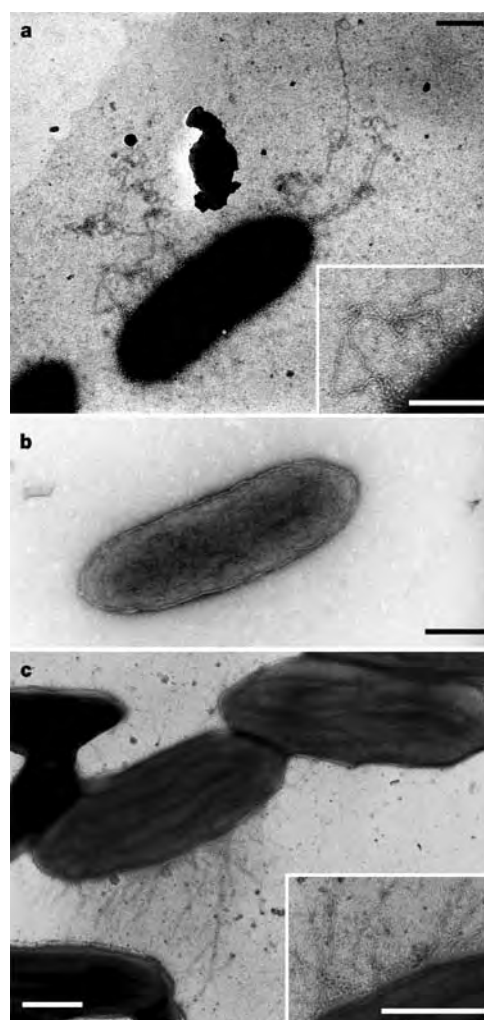


Figure 2 | Transmission electron microscopy analyses. Shown are cells of a wild-type strain (a), a *pilA*-deficient mutant strain (b) and a complemented mutant strain (c) of *G. sulfurreducens*. Cells were grown in medium with acetate and fumarate at 25 °C to induce the formation of pili, then negatively stained. Insets in a and c show details of pili produced by the wild-type and complemented mutant strains, respectively. Scale bars, 0.2 μ m.

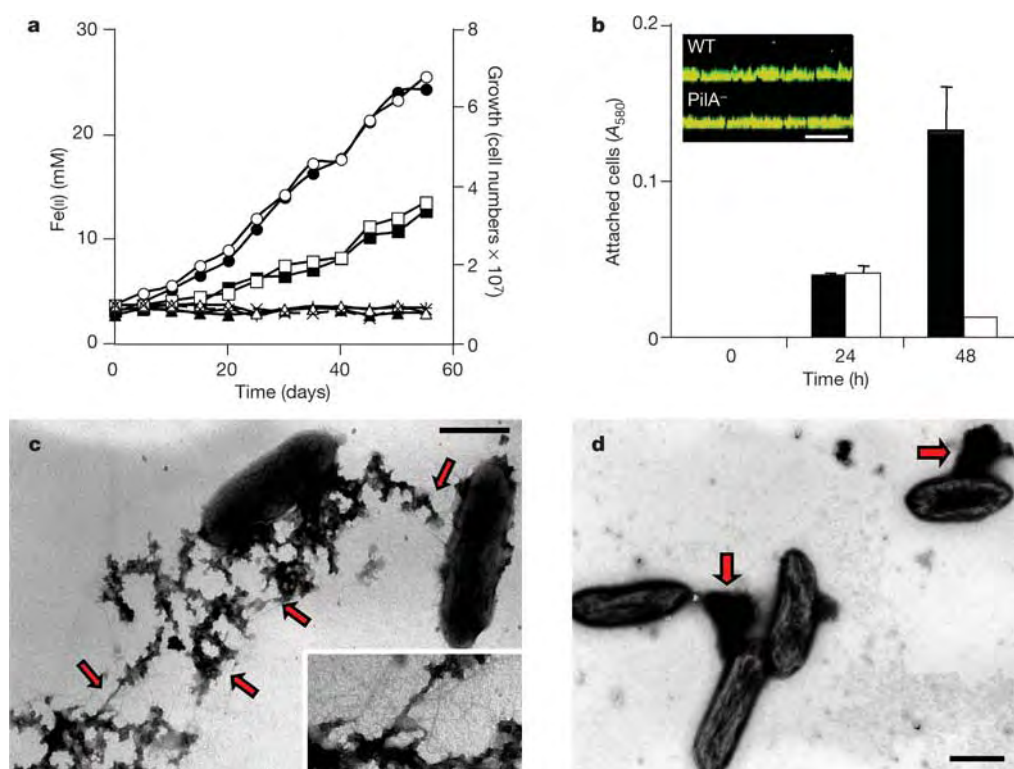


Figure 3 | Effect of a mutation in pilin production on reduction of Fe(III) oxide and attachment. **a**, Cells (open symbols) of the wild-type (circles), $\Delta pilA$ mutant (triangles) and complemented $\Delta pilA$ mutant (squares) strains and Fe(II) produced from Fe(III) reduction (filled symbols). Plus signs, Fe(II) in uninoculated medium; crosses, cells in uninoculated medium. **b**, Biomass of cells attached to Fe(III) oxide-coated coverslips over time; inset, confocal scanning laser microscopy of biomass attached in the first 24 h to the Fe(III) oxide, which is at the bottom of each image. Scale bar, 20 μ m. Solid bars, wild type (WT); open bars, $\Delta pilA$ mutant ($PilA^-$). Error bars show s.d. **c, d**, Transmission electron micrographs of fumarate-grown WT (**c**) and $PilA^-$ (**d**) cells amended with Fe(III) oxides (indicated by arrows). Scale bars, 0.5 μ m. Inset in (**c**) shows WT pili intertwined with Fe(III) oxides.

reducers^{1,3}. However, the outer-membrane cytochrome model for Fe(III) reduction has serious limitations. For example, whereas *Geobacter* and *Desulfuromonas* species in the *Geobacteraceae* family of Fe(III)-reducing microorganisms contain abundant *c*-type cytochromes, no *c*-type cytochromes could be detected in *Pelobacter* species⁴, which are phylogenetically intertwined with *Geobacter*

and *Desulfuromonas* species^{20,21}. Yet *Pelobacter* species—which, like *Geobacter* species, contain pili localized on one side of the cell—also are capable of reducing Fe(III) oxides⁴.

The conductive pili provide the opportunity to extend electron transfer capabilities well beyond the outer surface of the cells, which might be especially important in soils in which Fe(III) oxides exist as

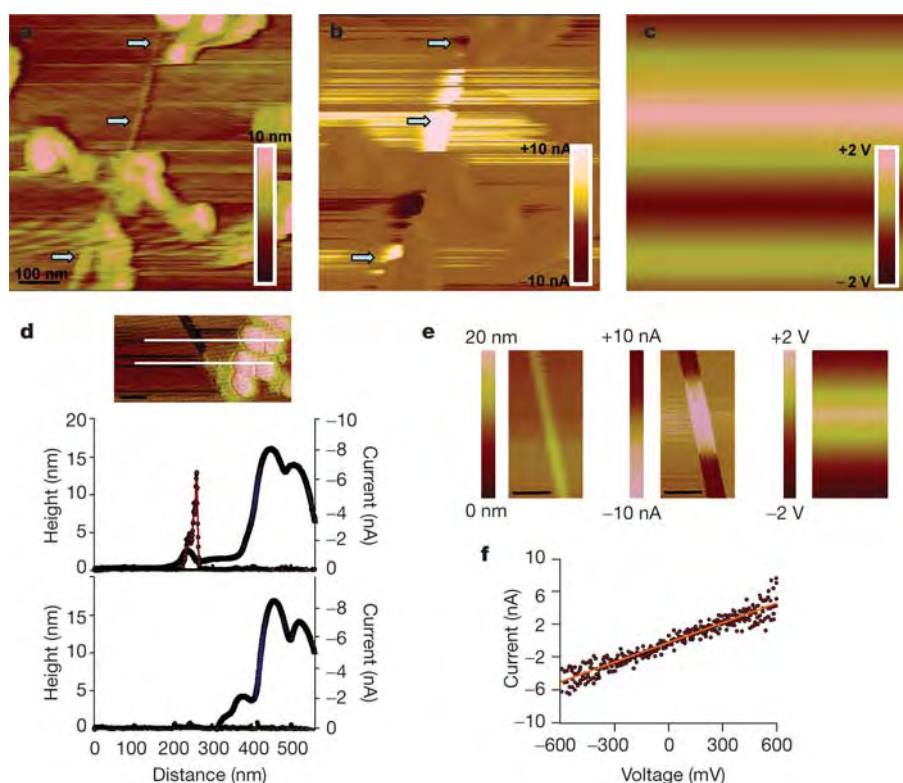


Figure 4 | Conducting-probe atomic force microscopy. **a**, Topography of a pilus (indicated by arrows) and non-pilus globular proteins. **b**, **c**, Current image (**b**) of the same field when a slow, triangular sweep bias voltage (**c**) was applied to the tip. **d**, Image showing both the topography and current of a pilus (top); the top and bottom white lines show the locations of the scans in the middle and bottom panels, respectively. Thick line (blue open circles where visible), height; red line, current. **e**, Results from disabling the slow axis to repeatedly scan horizontally across the same portion of a pilus filament. The apparent increased width of the pilus is an artefact of this form of scanning. **f**, Correspondence between current and applied voltage. Scale bars, 100 nm.

heterogeneously dispersed coatings on clays and other particulate matter. The pilus apparatus is anchored in the periplasm and outer membrane of Gram-negative cells, thus offering the possibility that pili accept electrons from periplasmic and/or outer membrane electron transfer proteins. These intermediary electron transfer proteins need not be the same in all organisms, which is consistent with the differences in cytochrome content and/or composition in different Fe(III) reducers¹. The likely function of the pili is to complete the circuit between these various intermediary electron carriers and the Fe(III) oxide.

In addition to serving as a conduit for electron transfer to Fe(III) oxides, pili could conceivably be involved in other electron transfer reactions. For example, pili of individual *Geobacter* cells are often intertwined, raising the possibility of cell-to-cell electron transfer through pili. These biologically produced nanowires might be useful in nanoelectronic applications^{22,23}, with the possibility of genetically modifying pilin structure and/or composition to generate nanowires with different functionalities.

METHODS

Bacterial strains and culture conditions. All *G. sulfurreducens* strains were isogenic with the wild-type strain PCA (ATCC 51573). A PilA[−] mutant strain was generated by replacement of the +61 to +159 coding region of the *pilA* gene (GSU1496) with a chloramphenicol cassette, as described previously⁹. The *pilA* mutation was complemented *in trans* by introducing plasmid pRG5-*pilA*, a pRG5 derivative²⁴ carrying a wild-type copy of the coding region of *pilA*.

Cells were routinely cultured at 30 or 25 °C under strictly anaerobic conditions in freshwater medium supplemented with acetate as electron donor, with fumarate, Fe(III)-citrate or poorly crystalline Fe(III) oxides (100 mM) as the electron acceptor²⁵, and in the presence of chloramphenicol (15 µg ml^{−1}) or spectinomycin (150–300 µg ml^{−1}) for cultures of the PilA[−] and pRG5-*pilA* strains, respectively. Rates of Fe(III) oxide reduction were determined by measuring the production of Fe(II), and growth was determined as cell counts of acridine-orange-stained cells. For attachment assays, cells were grown in the presence of coverslips coated with Fe(III) oxide; the attached biomass was determined with crystal violet (see Supplementary Information).

Conducting-probe atomic force microscopy analyses. Pili and other outer-surface proteins that were sheared from the cell surface (see Supplementary Information) were left to adsorb for 20 min on the surface of freshly cleaved, highly oriented pyrolytic graphite, fixed with 1% glutaraldehyde for 5 min, washed twice with double deionized water and blotted dry. Samples were examined with a Veeco Dimension 3100 AFM equipped with a Nanoscope IV controller and a SAM III signal access module to enable electrical interfacing with the tip. A gold-coated AFM tip (nominal spring constant 0.06 N m^{−1}; Veeco Inc.) was used for the imaging. The AFM was operated in contact mode with simultaneous tip–substrate conductivity mapping. While imaging, a slow ‘triangle-sweep’ bias voltage was applied to the tip in reference to the graphite surface by using a low-noise battery-powered ramping circuit. Current was measured with a DL Instruments 1211 current preamplifier.

Received 31 October 2004; accepted 21 April 2005.

1. Lovley, D. R., Holmes, D. E. & Nevin, K. P. in *Advances in Microbial Physiology* Vol. 49 (ed. Poole, R. K.) 219–286 (Elsevier Academic, London, 2004).
2. Lovley, D. R. Cleaning up with genomics: applying molecular biology to bioremediation. *Nature Rev. Microbiol.* **1**, 35–44 (2003).
3. Richardson, D. J. Bacterial respiration: a flexible process for a changing environment. *Microbiology* **146**, 551–571 (2000).
4. Lovley, D. R., Phillips, E. J., Lonergan, D. J. & Widman, P. K. Fe(III) and SO

reduction by *Pelobacter carbinolicus*. *Appl. Environ. Microbiol.* **61**, 2132–2138 (1995).

5. Nevin, K. P. & Lovley, D. R. Lack of production of electron-shuttling compounds or solubilization of Fe(III) during reduction of insoluble Fe(III) oxide by *Geobacter metallireducens*. *Appl. Environ. Microbiol.* **66**, 2248–2251 (2000).
6. Childers, S. E., Ciufo, S. & Lovley, D. R. *Geobacter metallireducens* accesses insoluble Fe(III) oxide by chemotaxis. *Nature* **416**, 767–769 (2002).
7. Nassif, X. et al. Type-4 pili and meningococcal adhesiveness. *Gene* **192**, 149–153 (1997).
8. Doig, P. et al. Role of pili in adhesion of *Pseudomonas aeruginosa* to human respiratory epithelial cells. *Infect. Immun.* **56**, 1641–1646 (1988).
9. Coppi, M. V., Leang, C., Sandler, S. J. & Lovley, D. R. Development of a genetic system for *Geobacter sulfurreducens*. *Appl. Environ. Microbiol.* **67**, 3180–3187 (2001).
10. Methé, B. A. et al. Genome of *Geobacter sulfurreducens*: metal reduction in subsurface environments. *Science* **302**, 1967–1969 (2003).
11. Sahu, S. N. et al. The bacterial adaptive response gene, *barA*, encodes a novel conserved histidine kinase regulatory switch for adaptation and modulation of metabolism in *Escherichia coli*. *Mol. Cell. Biochem.* **253**, 167–177 (2003).
12. Strom, M. S. & Lory, S. Structure-function and biogenesis of the type IV pili. *Annu. Rev. Microbiol.* **47**, 565–596 (1993).
13. Childers, S. E., Mehta, T., Ciufo, S. & Lovley, D. R. Abstracts 103rd General Meeting 361 (American Society for Microbiology, Washington DC, 2003).
14. Parge, H. E. et al. Structure of the fibre-forming protein pilin at 2.6 Å resolution. *Nature* **378**, 32–38 (1995).
15. Alm, R. A. & Mattick, J. S. Genes involved in the biogenesis and function of type-4 fimbriae in *Pseudomonas aeruginosa*. *Gene* **192**, 89–98 (1997).
16. Aho, E. L., Murphy, G. L. & Cannon, J. G. Distribution of specific DNA sequences among pathogenic and commensal *Neisseria* species. *Infect. Immun.* **55**, 1009–1013 (1987).
17. Wall, D. & Kaiser, D. Type IV pili and cell motility. *Mol. Microbiol.* **32**, 1–10 (1999).
18. Lovley, D. R., Coates, J. D., Blunt-Harris, E. L., Phillips, E. J. P. & Woodward, J. C. Humic substances as electron acceptors for microbial respiration. *Nature* **382**, 445–448 (1996).
19. Merz, A. J., So, M. & Sheetz, M. P. Pilus retraction powers bacterial twitching motility. *Nature* **407**, 98–102 (2000).
20. Holmes, D. E., Nevin, K. P. & Lovley, D. R. Comparison of 16S rRNA, *nifD*, *recA*, *gyrB*, *rpoB* and *fusA* genes within the family *Geobacteraceae* fam. nov. *Int. J. Syst. Evol. Microbiol.* **54**, 1591–1599 (2004).
21. Lonergan, D. J. et al. Phylogenetic analysis of dissimilatory Fe(III)-reducing bacteria. *J. Bacteriol.* **178**, 2402–2408 (1996).
22. Davis, J. J. Molecular bioelectronics. *Phil. Trans. R. Soc. Lond. A* **361**, 2807–2825 (2003).
23. Saylor, G. S., Simpson, M. L. & Cox, C. D. Emerging foundations: nano-engineering and bio-microelectronics for environmental biotechnology. *Curr. Opin. Microbiol.* **7**, 267–273 (2004).
24. Butler, J. E. et al. A single bifunctional enzyme for fumarate reduction and succinate oxidation in *Geobacter sulfurreducens* and *Geobacter metallireducens*. *J. Bacteriol.* (submitted).
25. Lovley, D. R. & Phillips, E. J. P. Novel mode of microbial energy metabolism: organic carbon oxidation coupled to dissimilatory reduction of iron or manganese. *Appl. Environ. Microbiol.* **54**, 1472–1480 (1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Russell and D. Bryant for helpful suggestions. This research was supported by grants to D.R.L. from the Department of Energy’s Genomics:GTL and NABIR programmes and DARPA, by a grant to M.T.T. from the National Science Foundation, and by a postdoctoral fellowship to G.R. from the Ministerio de Educación y Ciencia of Spain.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.R.L. (dlovley@microbio.umass.edu).

LETTERS

Invariant visual representation by single neurons in the human brain

R. Quian Quiroga^{1,2,†}, L. Reddy¹, G. Kreiman³, C. Koch¹ & I. Fried^{2,4}

It takes a fraction of a second to recognize a person or an object even when seen under strikingly different conditions. How such a robust, high-level representation is achieved by neurons in the human brain is still unclear^{1–6}. In monkeys, neurons in the upper stages of the ventral visual pathway respond to complex images such as faces and objects and show some degree of invariance to metric properties such as the stimulus size, position and viewing angle^{2,4,7–12}. We have previously shown that neurons in the human medial temporal lobe (MTL) fire selectively to images of faces, animals, objects or scenes^{13,14}. Here we report on a remarkable subset of MTL neurons that are selectively activated by strikingly different pictures of given individuals, landmarks or objects and in some cases even by letter strings with their names. These results suggest an invariant, sparse and explicit code, which might be important in the transformation of complex visual percepts into long-term and more abstract memories.

The subjects were eight patients with pharmacologically intractable epilepsy who had been implanted with depth electrodes to localize the focus of seizure onset. For each patient, the placement of the depth electrodes, in combination with micro-wires, was determined exclusively by clinical criteria¹³. We analysed responses of neurons from the hippocampus, amygdala, entorhinal cortex and parahippocampal gyrus to images shown on a laptop computer in 21 recording sessions. Stimuli were different pictures of individuals, animals, objects and landmark buildings presented for 1 s in pseudo-random order, six times each. An unpublished observation in our previous recordings was the sometimes surprising degree of invariance inherent in the neuron's (that is, unit's) firing behaviour. For example, in one case, a unit responded only to three completely different images of the ex-president Bill Clinton. Another unit (from a different patient) responded only to images of The Beatles, another one to cartoons from *The Simpson's* television series and another one to pictures of the basketball player Michael Jordan. This suggested that neurons might encode an abstract representation of an individual. We here ask whether MTL neurons can represent high-level information in an abstract manner characterized by invariance to the metric characteristics of the images. By invariance we mean that a given unit is activated mainly, albeit not necessarily uniquely, by different pictures of a given individual, landmark or object.

To investigate further this abstract representation, we introduced several modifications to optimize our recording and data processing conditions (see Supplementary Information) and we designed a paradigm to systematically search for and characterize such invariant neurons. In a first recording session, usually done early in the morning (screening session), a large number of images of famous persons, landmark buildings, animals and objects were shown. This set was complemented by images chosen after an interview with the

patient. The mean number of images in the screening session was 93.9 (range 71–114). The data were quickly analysed offline to determine the stimuli that elicited responses in at least one unit (see definition of response below). Subsequently, in later sessions (testing sessions) between three and eight variants of all the stimuli that had previously elicited a response were shown. If not enough stimuli elicited significant responses in the screening session, we chose those stimuli with the strongest responses. On average, 88.6 (range 70–110) different images showing distinct views of 14 individuals or objects (range 7–23) were used in the testing sessions. Single views of random stimuli (for example, famous and non-famous faces, houses, animals, etc) were also included. The total number of stimuli was determined by the time available with the patient (about 30 min on average). Because in our clinical set-up the recording conditions can sometimes change within a few hours, we always tried to perform the testing sessions shortly after the screening sessions in order to maximize the probability of recording from the same units. Unless explicitly stated otherwise, all the data reported in this study are from the testing sessions. To hold their attention, patients had to perform a simple task during all sessions (indicating with a key press whether a human face was present in the image). Performance was close to 100%.

We recorded from a total of 993 units (343 single units and 650 multi-units), with an average of 47.3 units per session (16.3 single units and 31.0 multi-units). Of these, 132 (14%; 64 single units and 68 multi-units) showed a statistically significant response to at least one picture. A response was considered significant if it was larger than the mean plus 5 standard deviations (s.d.) of the baseline and had at least two spikes in the post-stimulus time interval considered (300–1,000 ms). All these responses were highly selective: for the responsive units, an average of only 2.8% of the presented pictures (range: 0.9–22.8%) showed significant activations according to this criterion. This high selectivity was also present in the screening sessions, where only 3.1% of the pictures shown elicited responses (range: 0.9–18.0%). There was no significant difference between the relative number of responsive pictures obtained in the screening and testing sessions (*t*-test, *P* = 0.40). Responses started around 300 ms after stimulus onset and had mainly three non-exclusive patterns of activation (with about one-third of the cells having each type of response): the response disappeared with stimulus offset, 1 s after stimulus onset; it consisted of a rapid sequence of about 6 spikes (s.d. = 5) between 300 and 600 ms after stimulus onset; or it was prolonged and continued up to 1 s after stimulus offset. For this study, we calculated the responses in a time window between 300 and 1,000 ms after stimulus onset. In a few cases we also observed cells that responded selectively only after the image was removed from view (that is, after 1 s). These are not further analysed here.

¹Computation and Neural Systems, California Institute of Technology, Pasadena, California 91125, USA. ²Division of Neurosurgery and Neuropsychiatric Institute, University of California, Los Angeles (UCLA), California 90095, USA. ³Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.

⁴Functional Neurosurgery Unit, Tel-Aviv Medical Center and Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel. [†]Present address: Department of Engineering, University of Leicester, LE1 7RH, UK.

Figure 1a shows the responses of a single unit in the left posterior hippocampus to a selection of 30 out of the 87 pictures presented to the patient. None of the other pictures elicited a statistically significant response. This unit fired to all pictures of the actress Jennifer Aniston alone, but not (or only very weakly) to other famous and non-famous faces, landmarks, animals or objects. Interestingly, the unit did not respond to pictures of Jennifer Aniston together with the actor Brad Pitt (but see Supplementary Fig. 2). Pictures of Jennifer Aniston elicited an average of 4.85 spikes (s.d. = 3.59) between 300 and 600 ms after stimulus onset. Notably, this unit was nearly silent

during baseline (average of 0.02 spikes in a 700-ms pre-stimulus time window) and during the presentation of most other pictures (Fig. 1b). Figure 1b plots the median number of spikes (across trials) in the 300–1,000-ms post-stimulus interval for all 87 pictures shown to the patient. The histogram shows a marked differential response to pictures of Jennifer Aniston (red bars).

Next, we quantified the degree of invariance using a receiver operating characteristic (ROC) framework¹⁵. We considered as the hit rate (y axis) the relative number of responses to pictures of a specific individual, object, animal or landmark building, and as

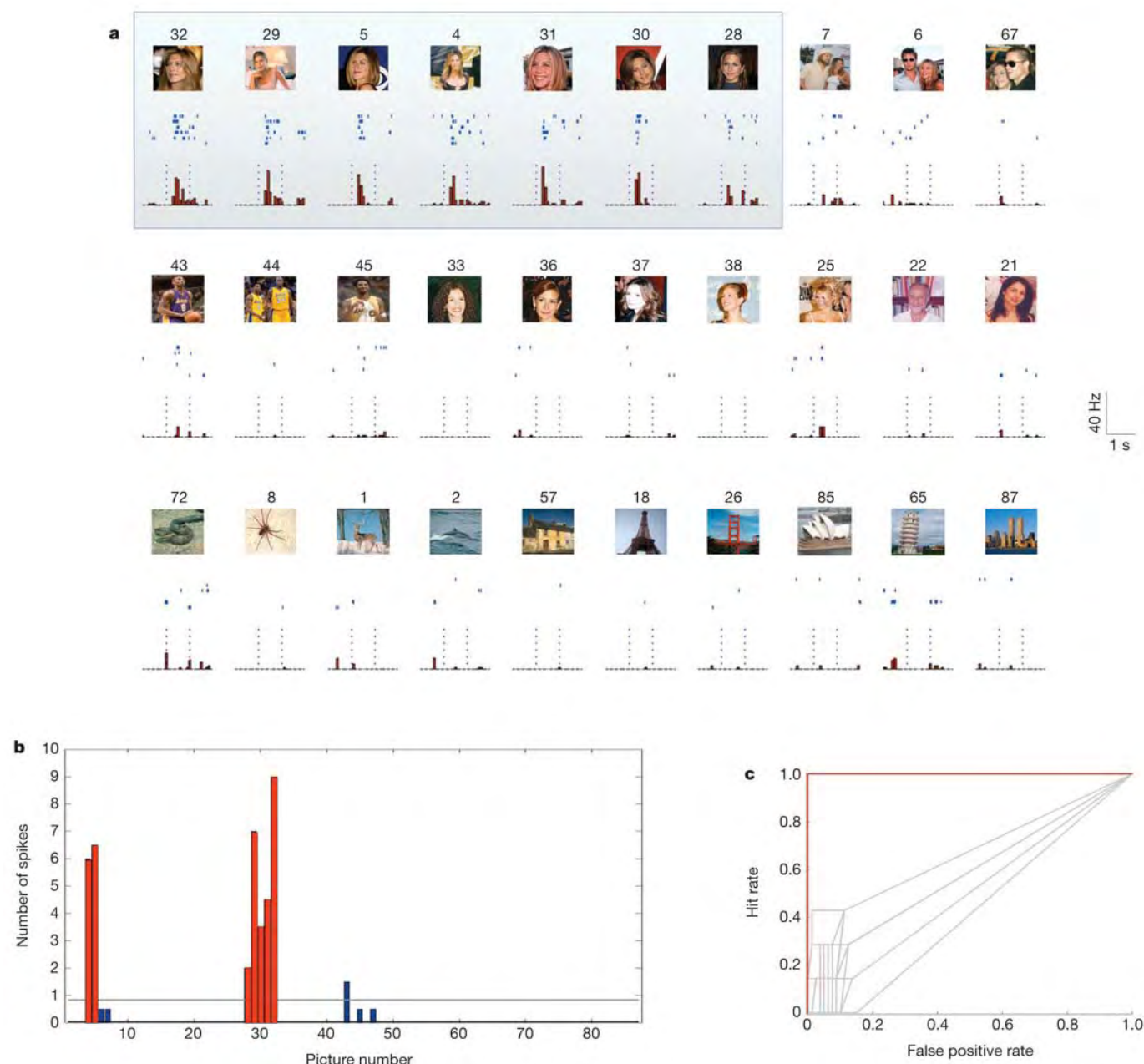


Figure 1 | A single unit in the left posterior hippocampus activated exclusively by different views of the actress Jennifer Aniston.

a, Responses to 30 of the 87 images are shown. There were no statistically significant responses to the other 57 pictures. For each picture, the corresponding raster plots (the order of trial number is from top to bottom) and post-stimulus time histograms are given. Vertical dashed lines indicate image onset and offset (1 s apart). Note that owing to insurmountable copyright problems, all original images were replaced in this and all subsequent figures by very similar ones (same subject, animal or building, similar pose, similar colour, line drawing, and so on). **b**, The median

responses to all pictures. The image numbers correspond to those in **a**. The two horizontal lines show the mean baseline activity (0.02 spikes) and the mean plus 5 s.d. (0.82 spikes). Pictures of Jennifer Aniston are denoted by red bars. **c**, The associated ROC curve (red trace) testing the hypothesis that the cell responded in an invariant manner to all seven photographs of Jennifer Aniston (hits) but not to other images (including photographs of Jennifer Aniston and Brad Pitt together; false positives). The grey lines correspond to the same ROC analysis for 99 surrogate sets of 7 randomly chosen pictures ($P < 0.01$). The area under the red curve is 1.00.

the false positive rate (x axis) the relative number of responses to other pictures. The ROC curve corresponds to the performance of a linear binary classifier for different values of a response threshold. Decreasing the threshold increases the probability of hits but also of false alarms. A cell responding to a large set of pictures of different individuals will have a ROC curve close to the diagonal (with an area under the curve of 0.5), whereas a cell that responds to all pictures of an individual but not to others will have a convex ROC curve far from the diagonal, with an area close to 1. In Fig. 1c we show the ROC curve for all seven pictures of Jennifer Aniston (red trace, with an area equal to 1). The grey lines show 99 ROC surrogate curves, testing invariance to randomly selected groups of pictures (see Methods). As expected, these curves are close to the diagonal, having an area of about 0.5. None of the 99 surrogate curves had an area equal or larger than the original ROC curve, implying that it is unlikely ($P < 0.01$)

that the responses to Jennifer Aniston were obtained by chance. A responsive unit was defined to have an invariant representation if the area under the ROC curve was larger than the area of the 99 surrogate curves.

Figure 2 shows another single unit located in the right anterior hippocampus of a different patient. This unit was selectively activated by pictures of the actress Halle Berry as well as by a drawing of her (but not by other drawings; for example, picture no. 87). This unit was also activated by several pictures of Halle Berry dressed as Catwoman, her character in a recent film, but not by other images of Catwoman that were not her (data not shown). Notably, the unit was selectively activated by the letter string 'Halle Berry'. Such an invariant pattern of activation goes beyond common visual features of the different stimuli. As with the previous unit, the responses were mainly localized between 300 and 600 ms after stimulus onset.

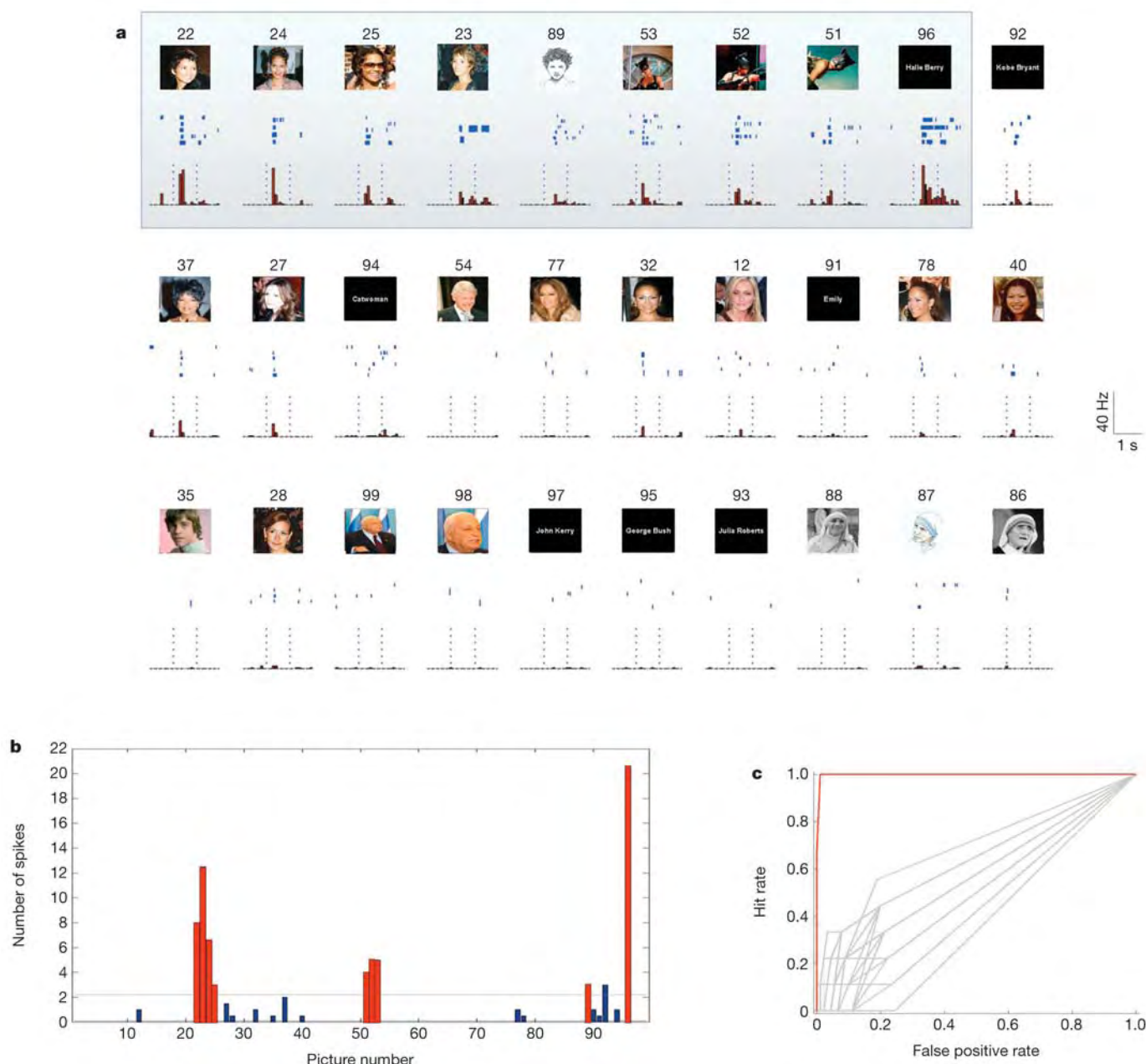


Figure 2 | A single unit in the right anterior hippocampus that responds to pictures of the actress Halle Berry (conventions as in Fig. 1).

a–c, Strikingly, this cell also responds to a drawing of her, to herself dressed as Catwoman (a recent movie in which she played the lead role) and to the

letter string 'Halle Berry' (picture no. 96). Such an invariant response cannot be attributed to common visual features of the stimuli. This unit also had a very low baseline firing rate (0.06 spikes). The area under the red curve in **c** is 0.99.

Figure 2c shows the ROC curve for the pictures of Halle Berry (red trace) and for 99 surrogates (grey lines). The area under the ROC curve was 0.99, larger than that of the surrogates.

Figure 3 illustrates a multi-unit in the left anterior hippocampus responding to pictures of the Sydney Opera House and the Baha'i Temple. Because the patient identified both landmark buildings as the Sydney Opera House, all these pictures were considered as a single landmark building for the ROC analysis. This unit also responded to the letter string 'Sydney Opera' (pictures no. 2 and 8) but not to other letter strings, such as 'Eiffel Tower' (picture no. 1). More examples of invariant responses are shown in the Supplementary Figs 2–11.

Out of the 132 responsive units, 51 (38.6%; 30 single units and 21 multi-units) showed invariance to a particular individual (38 units responding to Jennifer Aniston, Halle Berry, Julia Roberts, Kobe

Bryant, and so on), landmark building (6 units responding to the Tower of Pisa, the Baha'i Temple and the Sydney Opera House), animal (5 units responding to spiders, seals and horses) or object (2 units responding to specific food items), with $P < 0.01$ as defined above by means of the surrogate tests. A one-way analysis of variance (ANOVA) yielded similar results (see Methods). Eight of these units (two single units and six multi-units) responded to two different individuals (or to an individual and an object). Figure 4 presents the distribution of the areas under the ROC curves for all 51 units that showed an invariant representation to individuals or objects. The areas ranged from 0.76 to 1.00, with a median of 0.94. These units were located in the hippocampus (27 out of 60 responsive units; 45%), parahippocampal gyrus (11 out of 20 responsive units; 55%), amygdala (8 out of 30 responsive units; 27%) and entorhinal cortex

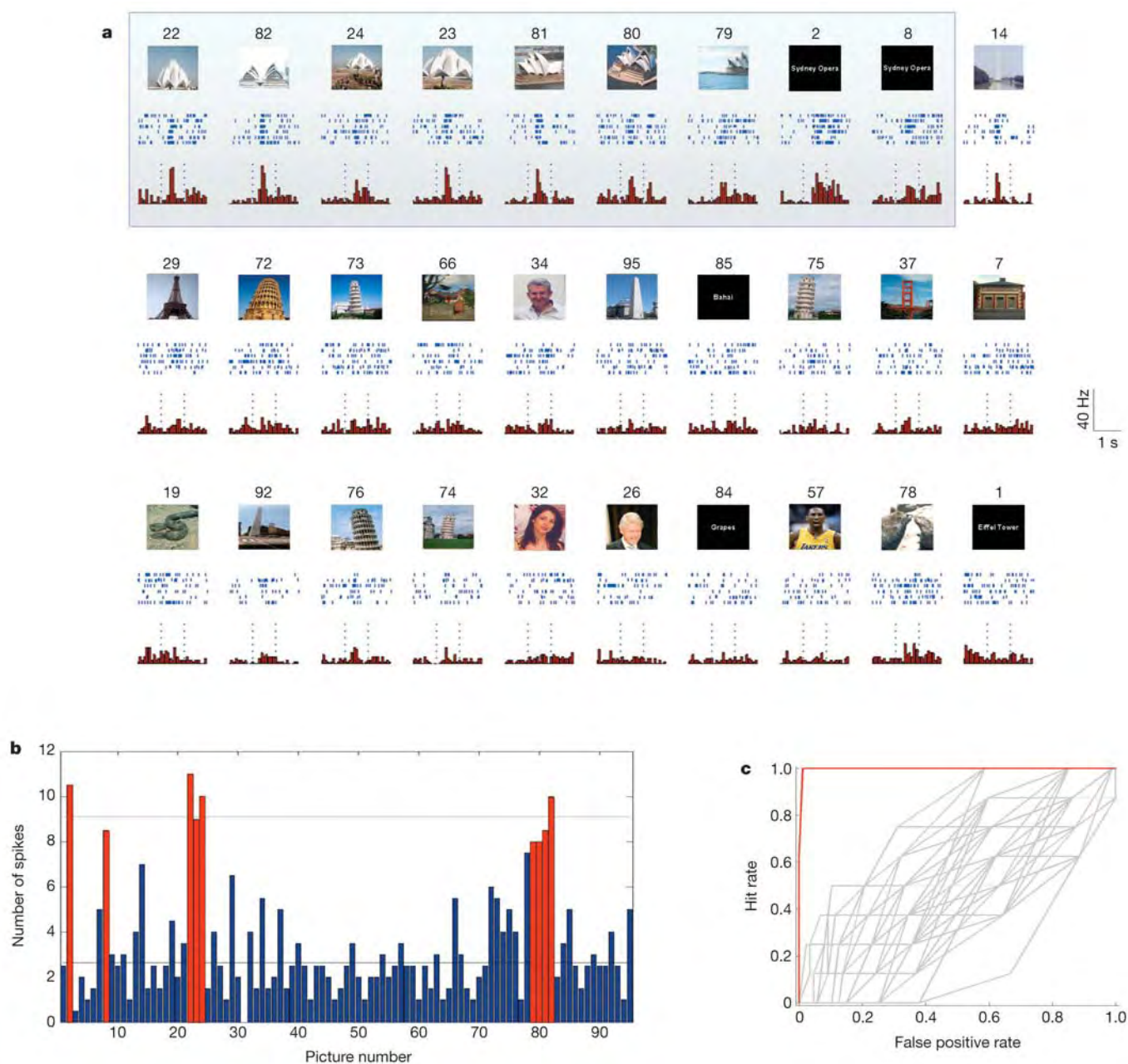


Figure 3 | A multi-unit in the left anterior hippocampus that responds to photographs of the Sydney Opera House and the Baha'i Temple (conventions as in Fig. 1). **a–c**, The patient identified all pictures of both of these buildings as the Sydney Opera, and we therefore considered them as a single landmark. This unit also responded to the presentation of the letter

string 'Sydney Opera' (pictures no. 2 and 8), but not to other strings, such as 'Eiffel Tower' (picture no. 1). In contrast to the previous two figures, this unit had a higher baseline firing rate (2.64 spikes). The area under the red curve in **c** is 0.97.

(5 out of 22 responsive units; 23%). There were no clear differences in the latencies and firing patterns among the different areas. However, more data are needed before making a conclusive claim about systematic differences between the various structures of the MTL.

As shown in Figs 2 and 3, one of the most extreme cases of an abstract representation is the one given by responses to pictures of a particular individual (or object) and to the presentation of the corresponding letter string with its name. In 18 of the 21 testing sessions we also tested responses to letter strings with the names of the individuals and objects. Eight of the 132 responsive units (6.1%) showed a selective response to an individual and its name (with no response to other names). Six of these were in the hippocampus, one was in the entorhinal cortex and one was in the amygdala.

These neuronal responses cannot be attributed to any particular movement artefact, because selective responses started around 300 ms after image onset, whereas key presses occurred at 1 s or later, and neuronal responses were very selective. About one-third of the responsive units had a response localized between 300 and 600 ms. This interval corresponds to the latency of event-related responses correlated with the recognition of 'oddball' stimuli in scalp electroencephalogram, namely, the P300 (ref. 16). Some studies argue for a generation of the P300 in the hippocampal formation and amygdala^{17,18}, consistent with our findings.

What are the common features that activate these neurons? Given the great diversity of distinct images of a single individual (pencil sketches, caricatures, letter strings, coloured photographs with different backgrounds) that these cells can selectively respond to, it is unlikely that this degree of invariance can be explained by a simple set of metric features common to these images. Indeed, our data are compatible with an abstract representation of the identity of the individual or object shown. The existence of such high-level visual responses in medial temporal lobe structures, usually considered to be involved in long-term memory formation and consolidation, should not be surprising given the following: (1) the known anatomical connections between the higher stages of the visual hierarchy in the ventral pathway and the MTL^{19,20}; (2) the well-characterized reactivity of the cortical stages feeding into the MTL to the sight of faces, objects, or spatial scenes (as ascertained using functional magnetic resonance imaging (fMRI) in humans^{21,22} and electrophysiology in monkeys^{2,4,7–11}); and (3) the observation that any visual percept that will be consciously remembered later on will have to be represented in the hippocampal system^{23–25}. This is true even though patients with bilateral loss of parts of the MTL do not, in general,

have a deficit in the perception of images²⁵. Neurons in the MTL might have a fundamental role in learning associations between abstract representations²⁶. Thus, our observed invariant responses probably arise from experiencing very different pictures, words or other visual stimuli in association with a given individual or object.

How neurons encode different percepts is one of the most intriguing questions in neuroscience. Two extreme hypotheses are schemes based on the explicit representations by highly selective (cardinal, gnostic or grandmother) neurons and schemes that rely on an implicit representation over a very broad and distributed population of neurons^{1–4,6}. In the latter case, recognition would require the simultaneous activation of a large number of cells and therefore we would expect each cell to respond to many pictures with similar basic features. This is in contrast to the sparse firing we observe, because most MTL cells do not respond to the great majority of images seen by the patient. Furthermore, cells signal a particular individual or object in an explicit manner²⁷, in the sense that the presence of the individual can, in principle, be reliably decoded from a very small number of neurons. We do not mean to imply the existence of single neurons coding uniquely for discrete percepts for several reasons: first, some of these units responded to pictures of more than one individual or object; second, given the limited duration of our recording sessions, we can only explore a tiny portion of stimulus space; and third, the fact that we can discover in this short time some images—such as photographs of Jennifer Aniston—that drive the cells suggests that each cell might represent more than one class of images. Yet, this subset of MTL cells is selectively activated by different views of individuals, landmarks, animals or objects. This is quite distinct from a completely distributed population code and suggests a sparse, explicit and invariant encoding of visual percepts in MTL. Such an abstract representation, in contrast to the metric representation in the early stages of the visual pathway, might be important in the storage of long-term memories. Other factors, including emotional responses towards some images, could conceivably influence the neuronal activity as well. The responses of these neurons are reminiscent of the behaviour of hippocampal place cells in rodents²⁸ that only fire if the animal moves through a particular spatial location, with the actual place field defined independently of sensory cues. Notably, place cells have been found recently in the human hippocampus as well²⁹. Both classes of neurons—place cells and the cells in the present study—have a very low baseline activity and respond in a highly selective manner. Future research might show that this similarity has functional implications, enabling mammals to encode behaviourally important features of the environment and to transition between them, either in physical space or in a more conceptual space¹³.

METHODS

The data in the present study come from 21 sessions in 8 patients with pharmacologically intractable epilepsy (eight right handed; 3 male; 17–47 years old). Extensive non-invasive monitoring did not yield concordant data corresponding to a single resectable epileptogenic focus. Therefore, the patients were implanted with chronic depth electrodes for 7–10 days to determine the seizure focus for possible surgical resection¹³. Here we report data from sites in the hippocampus, amygdala, entorhinal cortex and parahippocampal gyrus. All studies conformed to the guidelines of the Medical Institutional Review Board at UCLA. The electrode locations were based exclusively on clinical criteria and were verified by MRI or by computer tomography co-registered to preoperative MRI. Each electrode probe had a total of nine micro-wires at its end¹³, eight active recording channels and one reference. The differential signal from the micro-wires was amplified using a 64-channel Neuralynx system, filtered between 1 and 9,000 Hz. We computed the power spectrum for every unit after spike sorting. Units that showed evidence of line noise were excluded from subsequent analysis¹⁴. Signals were sampled at 28 kHz. Each recording session lasted about 30 min.

Subjects lay in bed, facing a laptop computer. Each image covered about 1.5° and was presented at the centre of the screen six times for 1 s. The order of the pictures was randomized. Subjects had to respond, after image offset, according to whether the picture contained a human face or something else by pressing the

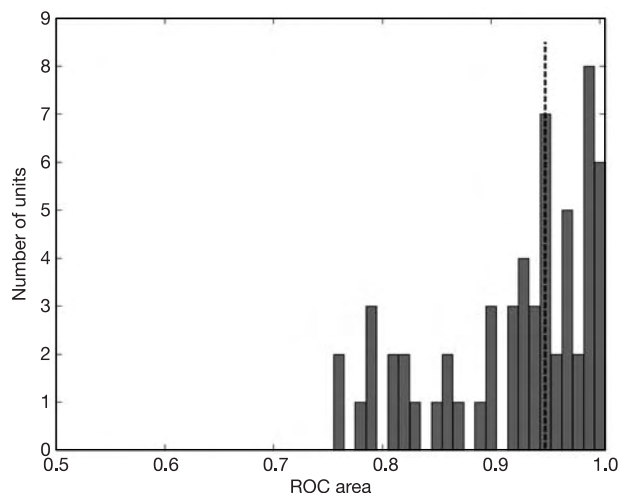


Figure 4 | Distribution of the area under the ROC curves for the 51 units (out of 132 responsive units) showing an invariant representation. Of these, 43 responded to a single individual or object and 8 to two individuals or objects. The dashed vertical line marks the median of the distribution (0.94).

'Y' and 'N' keys, respectively. This simple task, on which performance was virtually flawless, required them to attend to the pictures. After the experiments, patients gave feedback on whether they recognized the images or not. Pictures included famous and unknown individuals, animals, landmarks and objects. We tried to maximize the differences between pictures of the individuals (for example, different clothing, size, point of view, and so on). In 18 of the 21 sessions, we also presented letter strings with names of individuals or objects.

The data from the screening sessions were rapidly processed to identify responsive units and images. All pictures that elicited a response in the screening session were included in the later testing sessions. Three to eight different views of seven to twenty-three different individuals or objects were used in the testing sessions with a mean of 88.6 images per session (range 70–110). Spike detection and sorting was applied to the continuous recordings using a novel clustering algorithm³⁰ (see Supplementary Information). The response to a picture was defined as the median number of spikes across trials between 300 and 1,000 ms after stimulus onset. Baseline activity was the average spike count for all pictures between 1,000 and 300 ms before stimulus onset. A unit was considered responsive if the activity to at least one picture fulfilled two criteria: (1) the median number of spikes was larger than the average number of spikes for the baseline plus 5 s.d.; and (2) the median number of spikes was at least two.

The classification between single unit and multi-unit was done visually based on the following: (1) the spike shape and its variance; (2) the ratio between the spike peak value and the noise level; (3) the inter-spike interval distribution of each cluster; and (4) the presence of a refractory period for the single units (that is, less than 1% of spikes within less than 3 ms inter-spike interval).

Whenever a unit had a response to a given stimulus, we further analysed the responses to other pictures of the same individual or object by a ROC analysis. This tested whether cells responded selectively to pictures of a given individual. The hit rate (y axis) was defined as the number of responses to the individual divided by the total number of pictures of this individual. The false positive rate (x axis) was defined as the number of responses to the other pictures divided by the total number of other pictures. The ROC curve was obtained by gradually lowering the threshold of the responses (the median number of spikes in Figs 1b, 2b and 3b). Starting with a very high threshold (no hits, no false positives, lower left-hand corner in the ROC diagram), if a unit responds exclusively to an image of a particular individual or object, the ROC curve will show a steep increase when lowering the threshold (a hit rate of 1 and no false positives). If a unit responds to a random selection of pictures, it will have a similar relative number of hits and false positives and the ROC curve will fall along the diagonal. In the first case, for a highly invariant unit, the area under the ROC curve will be close to 1, whereas in the latter case it will be about 0.5. To evaluate the statistical significance, we created 99 surrogate curves for each responsive unit, testing the null hypothesis that the unit responded preferentially to n randomly chosen pictures (with n being the number of pictures of the individual for which invariance was tested). A unit was considered invariant to a certain individual or object if the area under the ROC curve was larger than the area of all of the 99 surrogates (that is, with a confidence of $P < 0.01$). Alternatively, the ROC analysis can be done with the single trial responses instead of the median responses across trials. Here, responses to the trials corresponding to any picture of the individual tested are considered as hits and responses to trials to other pictures as false positives. This trial-by-trial analysis led to very similar results, with 55 units of all 132 responsive units showing an invariant representation. A one-way ANOVA also yielded similar results. In particular, we tested whether the distribution of median firing rates for all responsive units showed a dependence on the factor identity (that is, the individual, landmark or object shown). The different views of each individual were the repeated measures. As with the ROC analysis, an ANOVA test was performed on all responsive units. Overall, the results were very similar to those obtained with the ROC analysis: of 132 responsive units, 49 had a significant effect for factor identity with $P < 0.01$, compared to 51 units showing an invariant representation with the ROC analysis. The ANOVA analysis, however, does not demonstrate that the invariant responses were very selective, whereas the ROC analysis explicitly tests the presence of an invariant as well as sparse representation.

Images were obtained from Corbis and Phototrazzi, with licensed rights to reproduce them in this paper and in the Supplementary Information.

Received 1 December 2004; accepted 3 February 2005.

- Barlow, H. Single units and sensation: a neuron doctrine for perception. *Perception* **1**, 371–394 (1972).
- Gross, C. G., Bender, D. B. & Rocha-Miranda, C. E. Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science* **166**, 1303–1306 (1969).
- Konorski, J. *Integrative Activity of the Brain* (Univ. Chicago Press, Chicago, 1967).
- Logothetis, N. K. & Sheinberg, D. L. Visual object recognition. *Annu. Rev. Neurosci.* **19**, 577–621 (1996).
- Riesenhuber, M. & Poggio, T. Neural mechanisms of object recognition. *Curr. Opin. Neurobiol.* **12**, 162–168 (2002).
- Young, M. P. & Yamane, S. Sparse population coding of faces in the inferior temporal cortex. *Science* **256**, 1327–1331 (1992).
- Logothetis, N. K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563 (1995).
- Logothetis, N. K. & Pauls, J. Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb. Cortex* **3**, 270–288 (1995).
- Perrett, D., Rolls, E. & Caan, W. Visual neurons responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* **47**, 329–342 (1982).
- Schwartz, E. L., Desimone, R., Albright, T. D. & Gross, C. G. Shape recognition and inferior temporal neurons. *Proc. Natl Acad. Sci. USA* **80**, 5776–5778 (1983).
- Tanaka, K. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* **19**, 109–139 (1996).
- Miyashita, Y. & Chang, H. S. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* **331**, 68–71 (1988).
- Fried, I., MacDonald, K. A. & Wilson, C. Single neuron activity in human hippocampus and amygdale during recognition of faces and objects. *Neuron* **18**, 753–765 (1997).
- Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neurosci.* **3**, 946–953 (2000).
- Macmillan, N. A. & Creelman, C. D. *Detection Theory: A User's Guide* (Cambridge Univ. Press, New York, 1991).
- Picton, T. The P300 wave of the human event-related potential. *J. Clin. Neurophysiol.* **9**, 456–479 (1992).
- Halgren, E., Marinkovic, K. & Chauvel, P. Generators of the late cognitive potentials in auditory and visual oddball tasks. *Electroencephalogr. Clin. Neurophysiol.* **106**, 156–164 (1998).
- McCarthy, G., Wood, C. C., Williamson, P. D. & Spencer, D. D. Task-dependent field potentials in human hippocampal formation. *J. Neurosci.* **9**, 4253–4268 (1989).
- Saleem, K. S. & Tanaka, K. Divergent projections from the anterior inferotemporal area TE to the perirhinal and entorhinal cortices in the macaque monkey. *J. Neurosci.* **16**, 4757–4775 (1996).
- Suzuki, W. A. Neuroanatomy of the monkey entorhinal, perirhinal and parahippocampal cortices: Organization of cortical inputs and interconnections with amygdale and striatum. *Seminars Neurosci.* **8**, 3–12 (1996).
- Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
- Haxby, J. V. et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
- Eichenbaum, H. A cortical-hippocampal system for declarative memory. *Nature Rev. Neurosci.* **1**, 41–50 (2000).
- Hampson, R. E., Pons, P. P., Stanford, T. R. & Deadwyler, S. A. Categorization in the monkey hippocampus: A possible mechanism for encoding information into memory. *Proc. Natl Acad. Sci. USA* **101**, 3184–3189 (2004).
- Squire, L. R., Stark, C. E. L. & Clark, R. E. The medial temporal lobe. *Annu. Rev. Neurosci.* **27**, 279–306 (2004).
- Mishashita, Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* **335**, 817–820 (1988).
- Koch, C. *The Quest for Consciousness: A Neurobiological Approach* (Roberts, Englewood, Colorado, 2004).
- Wilson, M. A. & McNaughton, B. L. Dynamics of the hippocampal ensemble code for space. *Science* **261**, 1055–1058 (1993).
- Ekstrom, A. D. et al. Cellular networks underlying human spatial navigation. *Nature* **425**, 184–187 (2003).
- Quiñero, R., Nadasdy, Z. & Ben-Shaul, Y. Unsupervised spike detection and sorting with wavelets and super-paramagnetic clustering. *Neural Comput.* **16**, 1661–1687 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank all patients for their participation; P. Sinha for drawing some faces; colleagues for providing pictures; I. Wainwright for administrative assistance; and E. Behnke, T. Fields, E. Ho, E. Isham, A. Kraskov, P. Steinmetz, I. Viskontas and C. Wilson for technical assistance. This work was supported by grants from the NINDS, NIMH, NSF, DARPA, the Office of Naval Research, the W.M. Keck Foundation Fund for Discovery in Basic Medical Research, a Whiteman fellowship (to G.K.), the Gordon Moore Foundation, the Sloan Foundation, and the Swartz Foundation for Computational Neuroscience.

Author Information Reprints and permissions information is available at ngp.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and request for materials should be addressed to R.Q.Q. (rodri@vis.caltech.edu).

LETTERS

An endocannabinoid mechanism for stress-induced analgesia

Andrea G. Hohmann¹, Richard L. Suplita¹, Nathan M. Bolton¹, Mark H. Neely¹, Darren Fegley², Regina Mangieri², Jocelyn F. Krey³, J. Michael Walker³, Philip V. Holmes¹, Jonathon D. Crystal¹, Andrea Duranti⁴, Andrea Tontini⁴, Marco Mor⁵, Giorgio Tarzia⁴ & Daniele Piomelli²

Acute stress suppresses pain by activating brain pathways that engage opioid or non-opioid mechanisms. Here we show that an opioid-independent form of this phenomenon, termed stress-induced analgesia¹, is mediated by the release of endogenous marijuana-like (cannabinoid) compounds in the brain. Blockade of cannabinoid CB₁ receptors in the periaqueductal grey matter of the midbrain prevents non-opioid stress-induced analgesia. In this region, stress elicits the rapid formation of two endogenous cannabinoids, the lipids 2-arachidonoylglycerol² (2-AG) and anandamide³. A newly developed inhibitor of the 2-AG-deactivating enzyme, monoacylglycerol lipase^{4,5}, selectively increases 2-AG concentrations and, when injected into the periaqueductal grey matter, enhances stress-induced analgesia in a CB₁-dependent manner. Inhibitors of the anandamide-deactivating enzyme fatty-acid amide hydrolase⁶, which selectively elevate anandamide concentrations, exert similar effects. Our results indicate that the coordinated release of 2-AG and anandamide in the periaqueductal grey matter might mediate opioid-independent stress-induced analgesia. These studies also identify monoacylglycerol lipase as a previously unrecognized therapeutic target.

Stress activates neural systems that inhibit pain sensation. This adaptive response, referred to as stress-induced analgesia (SIA), depends on the recruitment of brain pathways that project from the amygdala to the midbrain periaqueductal grey matter (PAG) and descend to the brainstem rostroventromedial medulla and dorsal horn of the spinal cord⁷. Endogenous opioid peptides have key functions in this process^{1,8}, but other as yet unidentified neurotransmitters are also known to be involved¹. We proposed that endocannabinoids might be implicated in stress analgesia for two reasons. First, agonists of CB₁ receptors—the predominant cannabinoid receptor subtype present in the brain^{9,10}—exert profound antinociceptive effects⁷ and suppress activity in nociceptive neurons^{11–14}. Second, CB₁ antagonists increase the activity of nociceptive rostroventromedial medulla neurons¹⁴ and enhance sensitivity to noxious stimuli¹⁵, indicating that an intrinsic endocannabinoid tone might regulate descending antinociceptive pathways⁷.

To study non-opioid SIA we delivered brief, continuous electric foot shock to rats and quantified their sensitivity to pain after stress by using the tail-flick test. As demonstrated previously^{1,16}, this stimulation protocol caused a profound antinociceptive effect that was not affected by intraperitoneal (i.p.) injection of the opiate antagonist naltrexone (14 mg kg⁻¹) (Fig. 1a). However, the response was almost abolished by administration of the CB₁ antagonist rimonabant (SR141617A, 5 mg kg⁻¹ i.p.) (Fig. 1a) or its analogue AM251 (5 mg kg⁻¹ i.p.) (Supplementary Fig. 1), but not by the CB₂

antagonist SR144528 (5 mg kg⁻¹ i.p.) (Fig. 1a). The effects of CB₁ antagonists cannot be attributed to changes in basal nociceptive threshold because, in the absence of the stressor, the drugs failed to alter tail-flick latencies (Supplementary Figs 1 and 2).

If CB₁ activation is required for the expression of non-opioid SIA, the latter should be lower in animals rendered tolerant to the

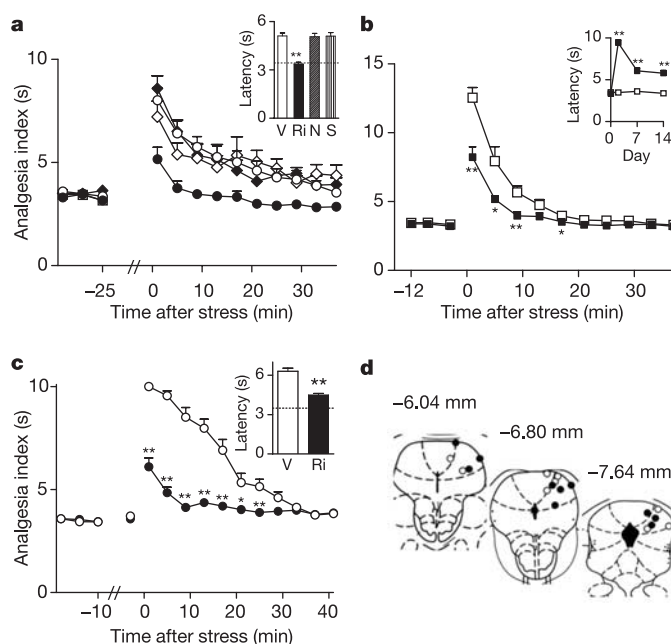


Figure 1 | CB₁ receptors mediate non-opioid stress-induced analgesia.

a, CB₁ antagonist rimonabant (Ri, filled circles) blocks stress antinociception. Opiate antagonist naltrexone (N, open diamonds) and CB₂ antagonist SR144528 (S, filled diamonds) have no effect. Open circles, vehicle (V). Inset: drug effects ($F_{3,29} = 5.99$, $P < 0.003$). The dotted line indicates the nociceptive threshold. Analgesia index was measured as the tail-flick latency. **b**, Stress antinociception is attenuated in WIN55212-2-tolerant rats (filled squares) compared with controls (open squares) ($F_{1,19} = 16.74$, $P < 0.0007$). Inset: non-stress cannabinoid antinociception is attenuated in WIN55212-2-tolerant rats ($F_{2,38} = 35.11$, $P < 0.0002$). **c**, Rimonabant in dorsolateral PAG suppresses stress antinociception ($F_{10,170} = 20.01$, $P < 0.0002$). Inset: drug effects ($F_{1,17} = 25.63$, $P < 0.0002$). **d**, PAG injection sites. Error bars, where visible, indicate s.e.m.; $n = 6–11$ per group. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ (analysis of variance, Fisher's PLSD test).

¹Neuroscience and Behavior Program, Department of Psychology, The University of Georgia, Athens, Georgia 30602-3013, USA. ²Department of Pharmacology and Center for Drug Discovery, University of California, Irvine, California 92697-4260, USA. ³Schrier Research Laboratory, Departments of Psychology and Neuroscience, Brown University, Providence, Rhode Island 02912, USA. ⁴Institute of Medicinal Chemistry, University of Urbino Carlo Bo, I-61029 Urbino, Italy. ⁵Pharmaceutical Department, University of Parma, I-43100 Parma, Italy.

antinociceptive effects of cannabinoids. Consistent with this prediction, rats chronically treated with the cannabinoid agonist WIN55212-2 (10 mg kg^{-1} i.p., once daily for 14 days) displayed, along with the expected blunting of acute CB_1 -dependent antinociception (Fig. 1b, inset), a marked decrease in stress antinociception (Fig. 1b). The possibility that this decrement might be due to changes in opioid tone is unlikely for two reasons: first, rats tolerant to WIN55212-2 showed no deficit in their antinociceptive response to morphine (2.5 mg kg^{-1} subcutaneously) (data not shown); and second, in accord with previous results¹⁶, rats tolerant to morphine (10 mg kg^{-1} subcutaneously once daily for 7 days) showed normal non-opioid stress antinociception (Supplementary Fig. 3).

The PAG serves key functions in both the descending control of pain^{7,17} and the antinociceptive actions of cannabinoid agonists¹⁸. We therefore asked whether blockade of CB_1 receptors in this structure could affect SIA. Rimobant (2 nmol) reduced stress antinociception when microinjected into the dorsolateral PAG (Fig. 1c, d), a structure linked to non-opioid stimulation-produced analgesia^{17,19}, but was inactive after injection into the lateral and ventrolateral PAG (Supplementary Fig. 4). The antagonist was also ineffective when administered into the lateral ventricle, indicating that its actions were not due to diffusion to distal sites (Supplementary Fig. 5). These results are consistent with the presence of CB_1 receptors throughout the dorsal midbrain^{9,20} and indicate that endocannabinoid release and/or intrinsic CB_1 receptor activity in the PAG might contribute to SIA.

To determine whether endocannabinoid release participates in this response, we measured anandamide and 2-AG concentrations in the dorsal midbrain of rats killed before (nonshock) or at various times after foot shock (Supplementary Fig. 6). Liquid chromatography/mass spectrometry (LC-MS) analyses revealed that midbrain 2-AG concentrations were markedly increased 2 min after shock termination and returned to baseline about 15 min later (Fig. 2a). This response preceded a sustained increase in anandamide concentration, which

peaked 7–15 min after foot shock (Fig. 2a). No such changes were observed in the occipital cortex (Fig. 2b), a brain region that contains CB_1 receptors⁹ but is not considered part of the SIA circuit.

The rapid accumulation of 2-AG in the midbrain after stress indicates that endocannabinoid release, rather than intrinsic CB_1 activity, might be responsible for SIA. If this is so, selective inhibitors of the 2-AG-hydrolysing enzyme monoacylglycerol lipase (MGL) should heighten the intrinsic actions of 2-AG and enhance its analgesic effects. The absence of selective MGL inhibitors prompted us to develop such an agent. To develop MGL-specific inhibitors we started from the assumption that similarities should exist between the substrate-binding site of MGL and that of the anandamide-hydrolysing enzyme fatty-acid amide hydrolase (FAAH)⁶: the fact that both hydrolases cleave arachidonic-acid derivatives indicates that their binding pockets might accommodate inhibitors of similar bulk and hydrophobicity. We therefore examined a collection of carbamate derivatives in which selective FAAH inhibition had been achieved by mimicking the flexible acyl chain of anandamide with the isosteric, but more rigid, biphenyl group (Fig. 3a)^{21,22}. This screening revealed that although *O*-biphenyl carbamates (Fig. 3a: 1, URB597; 2, URB524) inhibit the activity of FAAH but not that of MGL, *N*-biphenyl carbamates (Fig. 3a, 3, URB602) display an opposite selectivity (Fig. 3b, c). URB602 inhibited rat brain MGL with a half-maximal concentration (IC_{50}) of $28 \pm 4 \mu\text{M}$ (Fig. 3b) through a noncompetitive mechanism. Without URB602, the apparent Michaelis constant (K_m) of MGL for 2-AG was $24.0 \pm 1.7 \mu\text{M}$ and the maximum velocity (V_{max}) was $1814 \pm 51 \text{ nmol min per mg protein}$; with URB602, the K_m was $20.0 \pm 0.4 \mu\text{M}$ and the V_{max} was $541 \pm 20 \text{ nmol min per mg protein}$ ($n = 4$). When organotypic slice

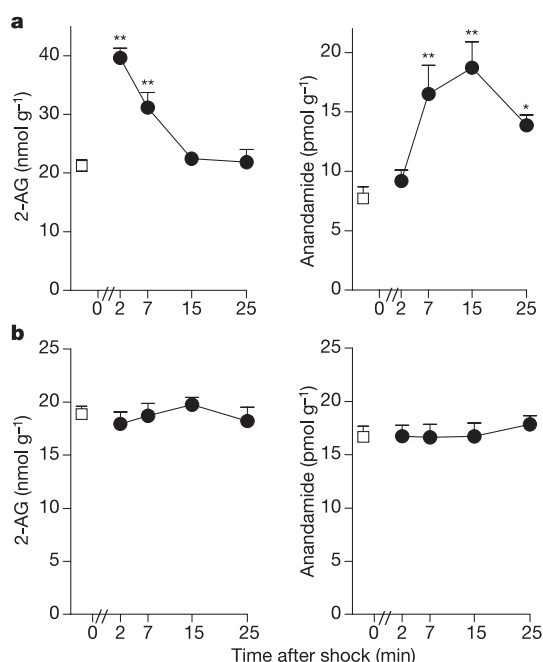


Figure 2 | Stress stimulates the formation of 2-AG and anandamide in dorsal midbrain. Non-stress (open squares) and post-stress (filled circles) concentrations of 2-AG and anandamide in dorsal midbrain samples containing the entire PAG (a) and in occipital cortex samples (b). Error bars indicate s.e.m.; $n = 10$ per group. Asterisk, $P < 0.05$ compared with non-stressed controls; two asterisks, $P < 0.01$.

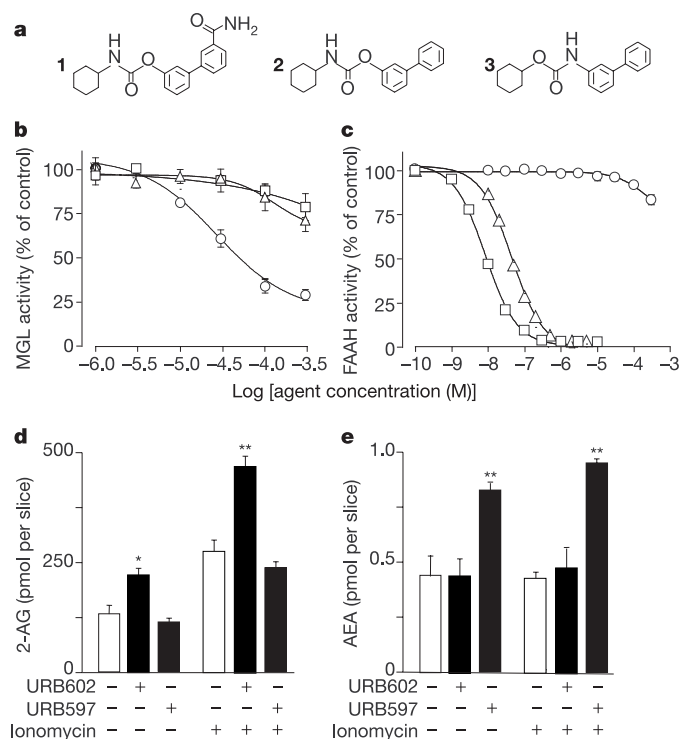


Figure 3 | URB602 is a selective MGL inhibitor. a, Structures of *O*-biphenyl-substituted FAAH inhibitors (1, URB597; 2, URB524) and the *N*-biphenyl-substituted MGL inhibitor URB602 (3). b, URB602 (circles) inhibits rat brain MGL activity, whereas URB597 (squares) and URB524 (triangles) have no such effect. c, URB602 does not affect rat brain FAAH activity, which is suppressed by URB597 and URB524. d, e, URB602 ($100 \mu\text{M}$) increases the concentration of 2-AG (d) but not of anandamide (e) in rat brain slice cultures. Effects of ionomycin ($2 \mu\text{M}$) and URB597 ($1 \mu\text{M}$) are also shown. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ versus control, *t*-test ($n = 4$). Error bars indicate s.e.m.

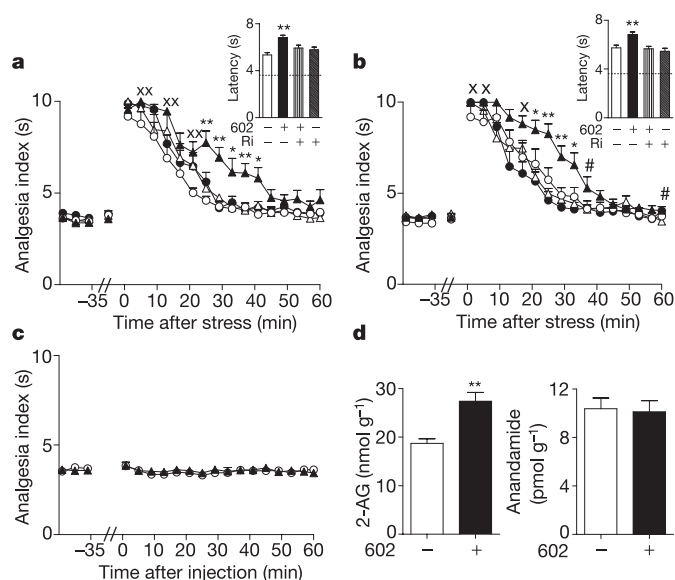


Figure 4 | The MGL inhibitor URB602 enhances non-opioid stress-induced analgesia. **a, b**, URB602 (602) increases stress antinociception when microinjected in dorsolateral PAG (**a**) and lateral/ventrolateral PAG (**b**), but does not cause antinociception in non-stressed rats (lateral/ventrolateral PAG) (**c**). Rimonabant blocks these effects. Open circles, vehicle; filled triangles, URB602; filled circles, rimonabant; open triangles, URB602/rimonabant. Analgesia index was measured as the tail-flick latency. Insets: drug effects (**a**, $F_{3,26} = 7.39$, $P < 0.002$; **b**, $F_{3,25} = 9.15$, $P < 0.0004$). Dotted lines indicate nociceptive thresholds. **d**, URB602 in the ventrolateral PAG increases the concentration of 2-AG, but not of anandamide, measured 25 min after shock. Error bars indicate s.e.m.; $n = 6-10$ per group. Asterisk, $P < 0.05$ compared with all groups; two asterisks, $P < 0.01$; cross, $P < 0.05$ compared with vehicle; two crosses, $P < 0.01$; hash, $P < 0.05$ compared with URB602/rimonabant; ANOVA, Fisher's PLSD post-hoc test.

cultures of rat forebrain were incubated with URB602 (100 μ M), both baseline and Ca^{2+} -ionophore-stimulated 2-AG concentrations were increased (Fig. 3d). In contrast, URB602 did not change anandamide content (Fig. 3e), which was markedly elevated by the FAAH inhibitor URB597 (ref. 21) at 1 μ M (Fig. 3e). Moreover,

URB602 did not affect the activities of lipid-metabolizing enzymes such as diacylglycerol lipase²³ and cyclooxygenase-2 (ref. 24) and did not significantly influence binding of [³H]WIN55212-2 to CB₁ or CB₂ receptors ($\text{IC}_{50} \geq 5 \mu\text{M}$) or [³⁵S]GTP- γ S to rat cerebellar membranes (half-maximal effective concentration (EC_{50}) $> 50 \mu\text{M}$) (Supplementary Table 1, and data not shown).

Because of its relatively low potency, URB602 is not suitable for systemic administration. Nevertheless, microinjections of the MGL inhibitor (0.1 nmol) into the dorsolateral PAG (Supplementary Fig. 7a) or lateral/ventrolateral PAG (Supplementary Fig. 7b) enhanced stress-induced antinociception (Fig. 4a, b). Basal nociceptive thresholds in non-shocked rats were unaffected (Fig. 4c; Supplementary Fig. 7c). This effect was probably due to the accumulation of 2-AG in the PAG, for three reasons. First, it was prevented by the simultaneous administration of rimonabant (0.2 nmol) (Fig. 4a, b). Second, it was mimicked by the non-selective MGL inhibitor methyl arachidonyl fluorophosphonate⁵ (2.6 nmol), whose effects also were blocked by rimonabant (Supplementary Fig. 9). Last, it was accompanied by an increase in midbrain 2-AG concentration: 25 min after foot shock, when the antinociceptive effect of URB602 was at its peak (Fig. 4a, b), 2-AG content was significantly higher in midbrain fragments of URB602-treated rats relative to vehicle-treated controls (Fig. 4d). Anandamide concentrations were identical in the two groups (Fig. 4d), further highlighting the selectivity of URB602 for MGL. These results indicate that URB602 is a selective MGL inhibitor that enhances stress antinociception.

To examine the possible role of anandamide in SIA, we administered the FAAH inhibitor URB597 (ref. 21) either by systemic (0.3 mg kg⁻¹, i.p.) (Fig. 5a) or local (0.1 nmol) (Fig. 5b; Supplementary Fig. 8) injection into the dorsolateral PAG. In both cases URB597 caused a potentiation of stress antinociception, which was prevented by rimonabant (1 mg kg⁻¹ i.p.; 0.2 nmol in the PAG) (Fig. 5a, b). The FAAH inhibitor did not modify basal nociceptive thresholds (Fig. 5a, b). Furthermore, administration of the anandamide transport inhibitor VDM11 (10 mg kg⁻¹ i.p.)²⁵ exerted similar effects, which also were blocked by rimonabant (2 mg kg⁻¹ i.p.) (Fig. 5c).

Our results indicate that the concerted release of 2-AG and anandamide in the PAG might mediate non-opioid SIA. The two endocannabinoids might act on local CB₁ receptors^{9,20,26} to regulate

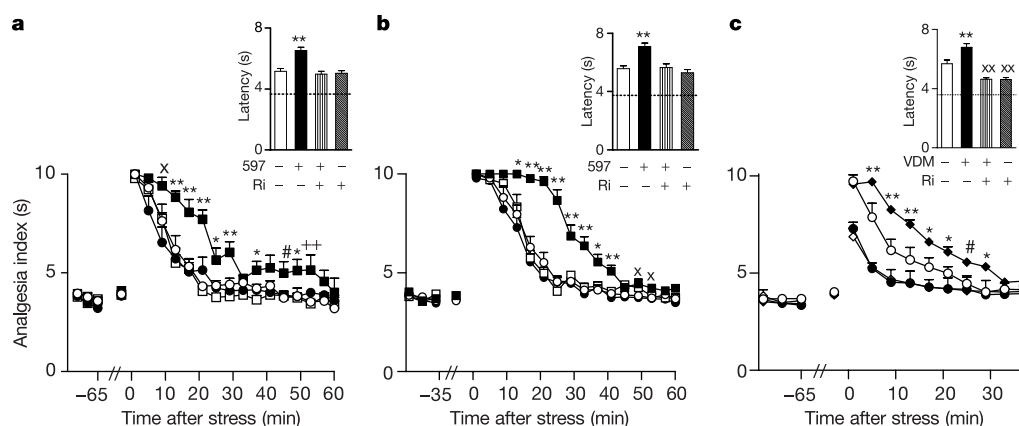


Figure 5 | Inhibitors of anandamide hydrolysis (URB597) and transport (VDM11) enhance non-opioid stress-induced analgesia. URB597 (filled squares) administered systemically (**a**) or in dorsolateral PAG (**b**) and VDM11 (filled diamonds) administered systemically (**c**) potentiate stress antinociception. Rimonabant blocks these effects after systemic administration (**a**, 1 mg kg⁻¹; **c**, 2 mg kg⁻¹) or administration in PAG (**b**). Vehicle, open circles; rimonabant, filled circles, URB597/rimonabant, open squares; VDM11/rimonabant, open diamonds. Analgesia index was

measured as the tail-flick latency. Insets: effects of URB597 (**a**, $F_{3,28} = 11.56$, $P < 0.0002$; **b**, $F_{3,24} = 33.69$, $P < 0.0002$) and VDM11 (**c**, $F_{3,26} = 21.76$, $P < 0.0002$). Error bars indicate s.e.m.; $n = 5-9$ per group. Asterisk, $P < 0.05$ compared with all groups; two asterisks, $P < 0.01$; cross, $P < 0.05$ compared with vehicle; two crosses, $P < 0.01$; hash, $P < 0.05$ compared with rimonabant (ANOVA, Fisher's PLSD post-hoc test). Dotted lines indicate nociceptive thresholds.

glutamate- and GABA-mediated transmission, ultimately disinhibiting descending pain control pathways. Three points are noteworthy. First, endocannabinoid-dependent stress antinociception is not affected by opioid antagonists or morphine tolerance, implying that it might not require opioid activity. The reverse may not be true, however, because mutant CB₁-null mice have reduced opioid-mediated responses to stress²⁷. Second, the residual antinociception observed in the presence of CB₁ antagonists leaves open the possibility that additional mediators of SIA remain to be discovered. Last, stress triggers the formation of both 2-AG and anandamide in the midbrain, but these two endocannabinoids are released with strikingly dissimilar time-courses. This observation underscores the existence of functional differences between these signalling molecules²⁸, pointing to the possibility that they might act in a coordinated manner to modulate temporally and/or spatially distinct processes in the PAG and other brain regions. The ability of both MGL and FAAH inhibitors to magnify endocannabinoid-dependent SIA also highlights the significance of these enzymes as previously undescribed targets for the treatment of pain and stress-related disorders.

METHODS

Chemicals. We synthesized URB597, URB524 and [²H₄]anandamide as described^{21,22,29,30}, and URB602 (biphenyl-3-yl carbamic acid cyclohexyl ester) by reacting diimidazole-1-ylmethanone with biphenyl-3-yl amine in acetonitrile in the presence of 4-dimethylaminopyridine and subsequently with cyclohexanol. Other chemicals were obtained and administered as described in Supplementary Methods.

Animals. We used adult male Sprague–Dawley rats for *in vivo* experiments and Wistar rats for enzyme assays and tissue cultures. All procedures were approved by the institutional animal care and use committee and followed guidelines of the International Association for the Study of Pain.

Brain slice cultures. We cultured brain slices from Wistar rats. Pups were killed on postnatal day 5 by decapitation after cryo-anaesthesia. Brains were removed and cut (0.4-mm-thick coronal slices) with a vibratome in a bath of ice-cold high-glucose DMEM (Gibco). Hemispheres were placed on Millicell culture inserts (Millipore) in six-well plates with serum-based culture medium (1.5 ml) composed of basal Eagle's medium with Earle's salts (100 ml), Earle's balanced salt solution (50 ml), heat-inactivated horse serum (50 ml), L-glutamine (0.2 mM, 1 ml) and 50% glucose (2 ml) (all from Gibco). Slices were maintained at 37°C with 5% CO₂ for 7 days before use.

Lipid extractions and LC–MS analyses. For *ex vivo* experiments, we habituated rats to the guillotine for at least 7 days before the experiment and killed them either before or at various times (2, 7, 15 and 25 min) after a 3-min foot shock ($n = 10$ per group). The brains were rapidly removed, dissected and stored frozen (–80°C) until lipid extraction. For *in vitro* experiments, we removed the medium of slice cultures and replaced it with DMEM (1 ml) containing URB602 (100 µM), URB597 (1 µM) or vehicle (0.1% dimethylsulphoxide) and incubated the slices at 25°C for 10 min. In some experiments, slices were incubated with ionomycin (2 µM) in DMEM for a further 15 min. Reactions were stopped and washed with ice-cold 50% methanol (1 ml). Slices were collected in the same medium (0.2 ml) and homogenized. Brain tissue (about 50 mg) and slice homogenates were suspended in methanol (2 ml) including ²H-containing internal standards (25 pmol). Lipids were extracted in methanol/chloroform/water (1:2:0.25). The organic phase was recovered, evaporated to dryness, reconstituted in chloroform/methanol (1:3, 80 µl) and subjected to LC–MS analysis as described³⁰.

Enzyme assays. We prepared cell fractions from Wistar rat brain homogenates, and assayed cytosol MGL activity and membrane FAAH activity with 2-monooleoyl[1,2,3-³H]glycerol (ARC; 20 Ci mmol^{–1}), and [ethanolamine-³H]anandamide (ARC; 60 Ci mmol^{–1}), respectively, as substrates^{4,21}.

Surgery and tolerance induction. We implanted stainless-steel guide cannulae in the PAG (dorsolateral or lateral/ventrolateral) or left lateral ventricle under pentobarbital/ketamine anaesthesia 3–7 days before testing. Placements of cannulae were verified in Nissl-stained sections or by post-mortem injection of fast green dye. Analyses were restricted to animals exhibiting dye spread throughout the ventricular system. The induction of tolerance to WIN55212-2 and morphine is described in Supplementary Methods.

Assessment of antinociception. We administered foot shock (0.9 mA, alternating current, for 3 min) to Sprague–Dawley rats with a Lafayette grid-shock apparatus. Withdrawal latencies in the radiant-heat tail-flick test^{13,17} were measured at 2-min intervals before (baseline) and after foot shock, and calculated for each subject in two-trial blocks. Removal of the tail from the

heat source terminated the application of thermal stimulation. Tail-flick latencies were monitored for 4 min immediately before exposure to the stressor to evaluate changes in nociceptive thresholds induced by pharmacological manipulations. Ceiling tail-flick latencies were 10 s except where noted. Tail-flick latencies, measured at baseline or before administration of the stressor, did not differ between groups in any study.

Data analyses. We analysed results with analysis of variance (ANOVA), repeated-measures ANOVA and Fisher's protected least-significant-difference post-hoc tests. $P < 0.05$ was considered significant.

Received 23 February; accepted 18 April 2005.

- Lewis, J. W., Cannon, J. T. & Liebeskind, J. C. Opioid and nonopioid mechanisms of stress analgesia. *Science* **208**, 623–625 (1980).
- Mechoulam, R. *et al.* Identification of an endogenous 2-monoglyceride, present in canine gut, that binds to cannabinoid receptors. *Biochem. Pharmacol.* **50**, 83–90 (1995).
- Devane, W. A. *et al.* Isolation and structure of a brain constituent that binds to the cannabinoid receptor. *Science* **258**, 1946–1949 (1992).
- Dinh, T. P. *et al.* Brain monoglyceride lipase participating in endocannabinoid inactivation. *Proc. Natl Acad. Sci. USA* **99**, 10819–10824 (2002).
- Dinh, T. P., Kathuria, S. & Piomelli, D. RNA interference suggests a primary role for monoacylglycerol lipase in the degradation of the endocannabinoid 2-arachidonoylglycerol. *Mol. Pharmacol.* **66**, 1260–1264 (2004).
- Cravatt, B. F. *et al.* Molecular characterization of an enzyme that degrades neuromodulatory fatty-acid amides. *Nature* **384**, 83–87 (1996).
- Walker, J. M. & Hohmann, A. G. In *Cannabinoids—Handbook of Experimental Pharmacology* (ed. Pertwee, R.) 509–554 (Springer, Berlin, 2005).
- Akil, H., Young, E., Walker, J. M. & Watson, S. J. The many possible roles of opioids and related peptides in stress-induced analgesia. *Ann. NY Acad. Sci.* **467**, 140–153 (1986).
- Herkenham, M. *et al.* Characterization and localization of cannabinoid receptors in rat brain: a quantitative *in vitro* autoradiographic study. *J. Neurosci.* **11**, 563–583 (1991).
- Zimmer, A., Zimmer, A. M., Hohmann, A. G., Herkenham, M. & Bonner, T. I. Increased mortality, hypoactivity, and hypoalgesia in cannabinoid CB₁ receptor knockout mice. *Proc. Natl Acad. Sci. USA* **96**, 5780–5785 (1999).
- Hohmann, A. G., Martin, W. J., Tsou, K. & Walker, J. M. Inhibition of noxious stimulus-evoked activity of spinal cord dorsal horn neurons by the cannabinoid WIN 55,212–2. *Life Sci.* **56**, 2111–2118 (1995).
- Hohmann, A. G., Tsou, K. & Walker, J. M. Cannabinoid suppression of noxious heat-evoked activity in wide dynamic range neurons in the lumbar dorsal horn of the rat. *J. Neurophysiol.* **81**, 575–583 (1999).
- Martin, W. J., Hohmann, A. G. & Walker, J. M. Suppression of noxious stimulus-evoked activity in the ventral posterolateral nucleus of the thalamus by a cannabinoid agonist: correlation between electrophysiological and antinociceptive effects. *J. Neurosci.* **16**, 6601–6611 (1996).
- Meng, I. D., Manning, B. H., Martin, W. J. & Fields, H. L. An analgesia circuit activated by cannabinoids. *Nature* **395**, 381–383 (1998).
- Calignano, A., La Rana, G., Giuffrida, A. & Piomelli, D. Control of pain initiation by endogenous cannabinoids. *Nature* **394**, 277–281 (1998).
- Terman, G. W., Lewis, J. W. & Liebeskind, J. C. Two opioid forms of stress analgesia: studies of tolerance and cross-tolerance. *Brain Res.* **368**, 101–106 (1986).
- Walker, J. M., Huang, S. M., Strangman, N. M., Tsou, K. & Sañudo-Peña, M. C. Pain modulation by release of the endogenous cannabinoid anandamide. *Proc. Natl Acad. Sci. USA* **96**, 12198–12203 (1999).
- Martin, W. J., Patrick, S. L., Coffin, P. O., Tsou, K. & Walker, J. M. An examination of the central sites of action of cannabinoid-induced antinociception in the rat. *Life Sci.* **56**, 2103–2109 (1995).
- Cannon, J. T., Prieto, G. J., Lee, A. & Liebeskind, J. C. Evidence for opioid and non-opioid forms of stimulation-produced analgesia in the rat. *Brain Res.* **243**, 315–321 (1982).
- Tsou, K., Brown, S., Sañudo-Peña, M. C., Mackie, K. & Walker, J. M. Immunohistochemical distribution of cannabinoid CB₁ receptors in the rat central nervous system. *Neuroscience* **83**, 393–411 (1998).
- Kathuria, S. *et al.* Modulation of anxiety through blockade of anandamide hydrolysis. *Nature Med.* **1**, 76–81 (2003).
- Mor, M. *et al.* Cyclohexylcarbamic acid 3'- or 4'-substituted biphenyl-3-yl esters as fatty acid amide hydrolase inhibitors: synthesis, quantitative structure–activity relationships, and molecular modeling studies. *J. Med. Chem.* **47**, 4998–5008 (2004).
- Stella, N., Schweitzer, P. & Piomelli, D. A second endogenous cannabinoid that modulates long-term potentiation. *Nature* **388**, 773–778 (1997).
- Kozak, K. R., Prusakiewicz, J. J. & Marnett, L. J. Oxidative metabolism of endocannabinoids by COX-2. *Curr. Pharm. Des.* **10**, 659–667 (2004).
- De Petrocellis, L., Bisogno, T., Davis, J. B., Pertwee, R. G. & Di Marzo, V. Overlap between the ligand recognition properties of the anandamide transporter and the VR1 vanilloid receptor: inhibitors of anandamide uptake with negligible capsaicin-like activity. *FEBS Lett.* **483**, 52–56 (2000).
- Vaughan, C. W., Connor, M., Bagley, E. E. & Christie, M. J. Actions of

- cannabinoids on membrane properties and synaptic transmission in rat periaqueductal gray neurons *in vitro*. *Mol. Pharmacol.* **57**, 288–295 (2000).
27. Valverde, O., Ledent, C., Beslot, F., Parmentier, M. & Roques, B. P. Reduction of stress-induced analgesia but not of exogenous opioid effects in mice lacking CB1 receptors. *Eur. J. Neurosci.* **12**, 533–539 (2000).
28. Piomelli, D. The molecular logic of endocannabinoid signalling. *Nature Rev. Neurosci.* **4**, 873–884 (2003).
29. Tarzia, G. et al. Design, synthesis, and structure–activity relationships of alkylcarbamic acid aryl esters, a new class of fatty acid amide hydrolase inhibitors. *J. Med. Chem.* **46**, 2352–2360 (2003).
30. Giuffrida, A., Rodríguez de Fonseca, F. & Piomelli, D. Quantification of bioactive acylethanolamides in rat plasma by electrospray mass spectrometry. *Anal. Biochem.* **280**, 87–93 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The assistance of the Centro di Calcolo at the University of Parma is gratefully acknowledged. This research was supported by grants from the National Institute on Drug Abuse (A.G.H., D.P.) and from the MIUR and the Universities of Parma and Urbino 'Carlo Bo'.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to A.G.H. (ahohmann@uga.edu) or D.P. (piomelli@uci.edu).

Early developmental arrest of mammalian limbs lacking *HoxA/HoxD* gene function

Marie Kmita¹, Basile Tarchini¹, Jozsef Zákány¹, Malcolm Logan², Clifford J. Tabin³ & Denis Duboule¹

Vertebrate *HoxA* and *HoxD* cluster genes are required for proper limb development^{1–3}. However, early lethality, compensation and redundancy have made a full assessment of their function difficult^{3–5}. Here we describe mice that are lacking all *Hoxa* and *Hoxd* functions in their forelimbs. We show that such limbs are arrested early in their developmental patterning and display severe truncations of distal elements, partly owing to the absence of *Sonic hedgehog* expression. These results indicate that the evolutionary recruitment of *Hox* gene function into growing appendages might have been crucial in implementing *hedgehog* signalling, subsequently leading to the distal extension of tetrapod appendages. Accordingly, these mutant limbs may be reminiscent of an ancestral trunk extension, related to that proposed for arthropods⁶.

Vertebrate limbs bud out of flank mesoderm through interactions with the overlying ectoderm. The subsequent outgrowth and patterning of skeletal elements require signals from both the apical ectodermal ridge (AER) and the zone of polarizing activity, a cohort of cells at the posterior margin of the bud. These cells express *Sonic hedgehog* (*Shh*)⁷, whose signalling promotes distal limb growth and patterning, notably through its effect on *Hox* genes belonging to both the *HoxA* and *HoxD* clusters. Before responding to *Shh* signalling, up to seven *Hox* genes are expressed in the early bud, with a restriction in the posterior part, where they may promote *Shh* transcription¹.

Functional analyses have highlighted the function of *Hox* genes in developing limbs. In particular, compound mutants revealed synergistic and redundant mechanisms, because phenotypic alterations were significantly more severe than merely additive. Although this raises problems in assigning gene-specific phenotypes, it indicates that *Hox* products might act quantitatively in both the production and the organization of the structure, a conclusion supported by the truncations observed in mice lacking one of group 13, 11 or 10 *Hox* genes^{2,3,5}. In contrast to the *HoxA* and *HoxD* clusters, *HoxB* and *HoxC* are unlikely to have a major function in forelimb development⁵, a conclusion based on expression analyses. Furthermore, normal limbs developed in the absence of these latter clusters^{8,9}. Consequently, to evaluate the extent of forelimb development in the absence of any relevant *Hox* function, we engineered a combined deletion of the *HoxA* and *HoxD* clusters.

Because loss of *Hoxa13* is embryonic lethal³, we floxed the *HoxA* cluster to generate tissue-specific deletions (Fig. 1a; *HoxA*^{lox}) and used *Prx1-Cre* mice, where recombination occurs from early limb bud stage onwards¹⁰. Mice homozygous for a conditional *HoxA* deletion (*HoxA*^{c/c}, referred to hereafter as *A*^{c/c}) were viable but infertile. We assessed the efficiency of recombination through the disappearance of *Hoxa* transcripts from developing *A*^{c/c} limbs (Fig. 1b). Although some *Hoxa13* and *Hoxa11* transcripts were still detectable in early buds, by late in day 11 *Hoxa* RNAs were no longer

seen in forelimbs (Fig. 1b). To decrease residual *Hoxa* transcripts further, we generated embryos *trans*-heterozygous for a full *HoxA* deletion (*HoxA*[–], or *A*[–]) and the conditional allele. Such embryos (*A*^{c/–}) had no trace of *Hoxa* transcripts in forelimb buds (Fig. 1b). Because *Prx1-Cre* was less efficient in hindlimbs (Fig. 1), only mutant forelimbs are documented.

Forelimb skeletons of *HoxA*^{lox/–}; *Prx1-Cre* adult mice (*A*^{c/–}) showed abnormal zeugopods (forearms) and autopods (hands; Fig. 1c). The thumb was absent and all digits were reduced in size. The zeugopod defects seemed stronger than the mere inactivation of *Hoxa11*, indicating that *Hoxa10* also participates in patterning this structure⁴. We verified that the *Prx1-Cre* transgene completely deleted *HoxA* before the time at which it became functional by generating compound mutants, in which *Hoxd11* was concomitantly inactivated. *A*^{c/–}; *Hoxd11*^{–/–} mice displayed markedly reduced zeugopods (Fig. 1c), comparable to those of *Hoxa11*; *Hoxd11* mutant

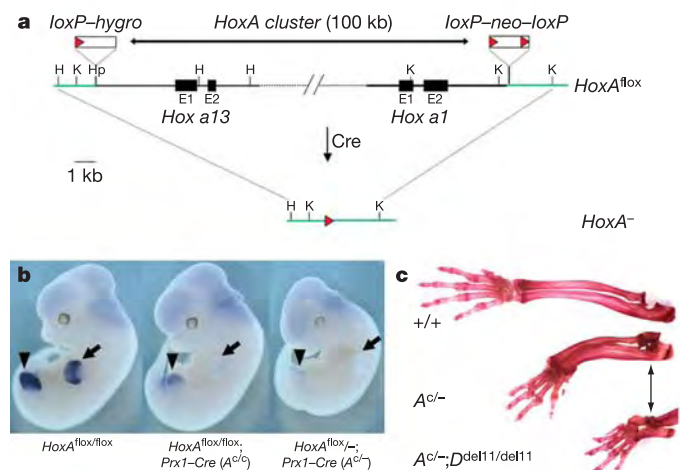


Figure 1 | Conditional deletion of the *HoxA* cluster. **a**, The cluster, flanked by *loxP* sites, was selectively deleted in limb buds. **b**, Control of deletion using a *Hoxa13* RNA probe. Floxed mice express *Hoxa13* in developing hindlimb (arrowhead) and forelimb (arrow). After deletion, *Hoxa13* expression in forelimbs is virtually undetectable, whereas transcripts are weakly detected in hindlimb buds. In E10 embryos carrying one floxed copy and one deleted allele (*A*^{c/–}), transcripts are undetectable in forelimbs (arrow) and barely scored in hindlimb buds (right). **c**, Adult forelimb of the *HoxA* conditional deletion (middle), showing a reduction in the length of bones and ill-formed radius and ulna. When combined with a *Hoxd11* inactivation (bottom), both radius and ulna were further reduced, resembling the double *Hoxd11*; *Hoxa11* mutant phenotype², indicating that *HoxA* function was absent from developing forelimbs.

¹Department of Zoology and Animal Biology and National Research Centre 'Frontiers in Genetics', University of Geneva, Sciences III, Quai Ernest Ansermet 30, 1211 Geneva 4, Switzerland. ²Division of Developmental Biology, National Institute for Medical Research, Mill Hill, London NW7 1AA, UK. ³Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

animals², showing that the conditional *HoxA* deletion occurred efficiently.

We combined a full deletion of *HoxD* ($D^{-/-}$)¹¹ together with one floxed and one deleted *HoxA* allele and *Prx1-Cre* ($HoxD^{-/-}$; $HoxA^{c/-}$). Double homozygous mutants died soon after birth; we therefore analysed fetal and newborn animals. Forelimbs deficient for *HoxA/HoxD* function were drastically truncated (Fig. 2b). At fetal stages, a single cartilage model was observed, articulating with the scapula. This cartilage element, bent in the middle, had a Y-like shape distally. We interpret this as a truncated humerus, bent distally and followed by a bifurcation prefiguring the formation of a zeugopod (Figs 2 and 3a–c). At birth, this cartilage was ossified proximally, whereas structure after bifurcation remained cartilaginous (Fig. 2b). Small and atypical distal rays were scored, which were identified not as proper digits but as distal derivatives of an abnormal cartilaginous plate (Fig. 3). Altogether, mutant forelimbs seemed delayed in their development, as though patterning had been arrested at an early stage. These skeletal elements are probably not produced through the persistence of undetectable *Hoxa* transcripts, because the phenotype of mice carrying a conditional *Tbx5* allele¹² indicates that the *Prx1-Cre* transgene might be active early on and throughout the proximal-to-distal extent of the limb bud. In addition, mice mutant for proximally expressed *Hox* genes^{13,14} have defects in the humerus, but not in its most proximal part, further indicating that the skeletal elements present in the mutant forelimbs of our mice might develop independently of any *Hoxa/Hoxd* function.

We compared this phenotype with fetuses lacking all relevant *Hox* gene function except that of *Hoxd13* by bringing the *HoxA* deletion over a *HoxD* deletion, removing *Hoxd9* to *Hoxd12* ($HoxD^{del9-12}$; B.T. and D.D., unpublished work). Proximal forelimb elements of such animals were comparable to those of $A^{c/-};D^{-/-}$ mutants, yet they displayed clear digits as early as embryonic day 15.5 (E15.5) (Fig. 3a). At birth, it was evident in these animals that the Y-shaped part was indeed a radius/ulna primordium, which generated two separate cartilages articulating with the humerus (Fig. 3d). Such individualized elements articulating with the humerus were not observed in forelimbs lacking all *Hoxa* and *Hoxd* genes. Comparison of both mutants confirmed that the atypical distal rays in $A^{c/-};D^{-/-}$ animals were not digits. They might reflect an intrinsic property of mutant distal mesenchyme in producing fragmented condensations. The fact that several digits normally derive from cells expressing *Shh*¹⁵, added to the absence of *Shh* expression in the *Hox*-deficient forelimbs we report here (see below), supports this conclusion.

The early developmental arrest of $A^{c/-};D^{-/-}$ forelimb buds was confirmed by the expression of *Meis1* and *dHand*. *Meis1* transcripts

normally mark the proximal half of the bud in E9.5–10.5 fetuses. Expression is subsequently maintained proximally as development proceeds (Fig. 3g). In mutant buds, *Meis1* expression at embryonic day 12.5 resembled the wild-type pattern at day 10 (Fig. 3f). Similarly, expression of *dHand* in E11.5 mutant limbs was identical to that of E10–10.5 wild-type buds (Fig. 3h, i) with a distal part devoid of transcript, whereas *dHand* expression in wild-type limbs showed a robust distal domain from E11 onwards (Fig. 3i). These results indicated that the early developmental programme was initiated but arrested soon afterwards.

Critical of this early-to-late transition is the activation of *Shh* transcription in the posterior limb bud. Forelimbs lacking *Shh* display severe distal and posterior agenesis, involving both the autopod and zeugopod, the humerus being less affected^{16–18}. We looked at *Shh* expression and observed an almost complete downregulation in conditional double-mutant forelimbs ($A^{c/-};D^{-/-}$) (Fig. 4a–e), with only few cells weakly positive (Fig. 4e). However, a single copy of either *HoxD* or *HoxA* was enough to trigger *Shh* transcription at a level similar to that in the wild type (Fig. 4b). To investigate further the requirement of *Hox* function for *Shh* transcription, we analysed embryos deficient for both clusters obtained by means of *trans*-heterozygous crosses ($A^{-/-};D^{-/-}$), before embryonic lethality. Two such embryos were obtained (out of 577) and *Shh* transcription was undetectable in the bud (Fig. 4, compare k to i and j), whereas other sites showed normal expression levels (Fig. 4, compare h to f, g).

These results indicate that the early expression of *Hox* genes in developing limbs is mandatory for *Shh* transcription to proceed. Forelimbs that lack *Shh* but have functional *Hox* clusters are less severely affected than is shown here. In particular, the radius and a distal element are produced, whereas they are absent from *Hox* mutant limbs. This difference reflects the persistence of some *Hox* transcripts derived from the budding stage in *Shh* mutant limbs^{16,19}. Downregulation of *Shh* signalling was further confirmed by the analysis of *Gli3* and *Fgf8* expression in double-mutant forelimbs. *Gli3* was expressed throughout the bud, unlike in control animals but as in *Shh* mutant mice. Similarly, *Fgf8* expression in the AER was abnormal and mostly absent from the distal parts (Fig. 4l, m).

Forelimbs of *HoxA/HoxD* mutants are therefore arrested in their development, at about the time at which SHH starts to accumulate posteriorly. It is difficult to infer from these results whether *Hox* genes act directly on *Shh* transcription and/or maintenance, or whether they sustain the growth of the bud by ensuring proper AER function. Nevertheless, the fact that early ectopic *Hox* expression can trigger *Shh* transcription^{1,20} argues in favour of the former

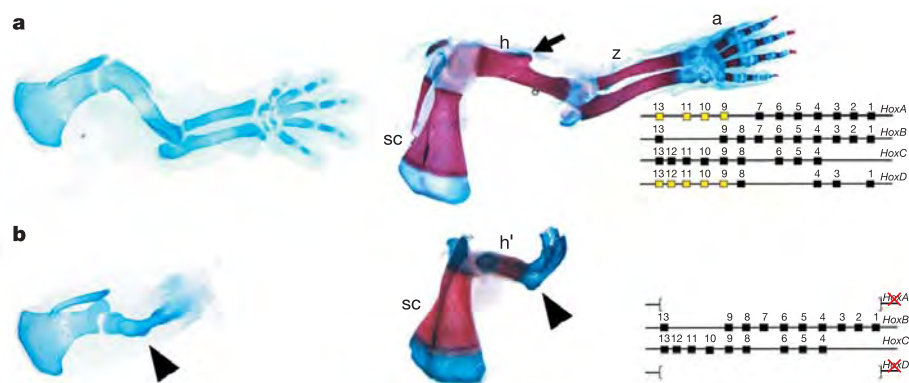


Figure 2 | *HoxA/HoxD* double-mutant forelimbs. Wild type (a) and mutant ($A^{c/-};D^{-/-}$) (b) cartilage and skeletal patterns in E15 (left) and newborn (right) specimens. Mutant animals (b) show a severe truncation of a single cartilage model (left), which is bent distally (arrowhead) and adopts a Y-shaped aspect. The scapula (sc) is apparently normal. After ossification

occurred, a single bone rod is observed proximally, articulating with the scapula. This truncated humerus (h') is extended, after the bend but without articulation, by a piece of cartilage fragmented at its distal tip. h, humerus; z, zeugopod; a, autopod.

alternative. This raises the possibility that the limb distal extension, in the course of tetrapod evolution, used the capacity of *Hox* gene function to trigger *Shh* transcription, which in turn antagonized the repressive effect of *Gli3* (refs 21,22). This required AER signalling to be fully functional, and defects in—or the absence of—either one of these components could prevent distal extension. In this view, the situation in teleost fishes, in which both the early *Hox* and the *Shh* patterns are present but no extension is observed, probably reflects a derived situation generated by modification in AER signalling after the ectodermal folding process in the fin bud²³.

Accordingly, the truncated proximal skeleton shown here might reflect an ancestral trunk extension, before the genetic machinery was co-opted to develop more distal limb pieces. Strikingly, most of these skeletal elements develop from cells expressing *Meis1*, a gene orthologous to the arthropod *homothorax* gene (*hth*). In *Drosophila*, *hth*, in

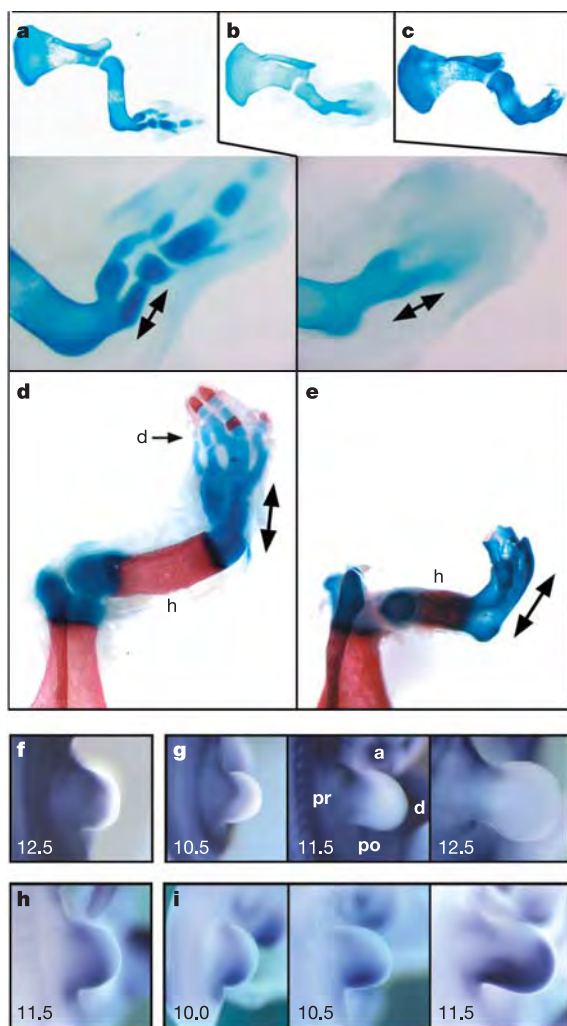


Figure 3 | Developmental arrest in forelimbs lacking *Hox* function. **a–c**, Cartilage patterns in E15.5 (**b**) and E17.5 (**c**) fetuses lacking *HoxA/HoxD* function in forelimbs ($A^{c/-};D^{-/-}$), or with *Hoxd13* function only ($A^{c/-};D^{\text{del9-12/del9-12}}$) in an E15.5 fetus (**a**). *Hoxd13* alone triggers the appearance of digits, distal to the Y-shaped structure (double arrow), absent from the double mutant (**b**). **d, e**, In postnatal skeletons (**d**), these condensations generated independent elements, remnants of radius and ulna, whereas the *HoxA/HoxD* mutant displayed a single cartilage plate (**e**, double arrow), with atypical rays. **f, g**, Expression of *Meis1* in mutant ($A^{c/-};D^{-/-}$) (**f**) and wild-type (**g**) buds. The mutant pattern at E12.5 resembles that of wild-type buds at E10.5, indicating a developmental arrest before E10. **h, i**, This is confirmed by *dHand* expression, which in E11.5 mutant ($A^{c/-};D^{-/-}$) buds (**h**) is expressed similarly to that in the wild type (**i**) at E10–E10.5. Abbreviations: a, anterior; po, posterior; pr, proximal; d, distal.

combination with *exd*, specifies an ancient limb proximal piece, whereas more distal structures emerge along with *Hedgehog* signalling^{24,25}. In vertebrates, limb bud expression of *Meis* represses *Shh* transcription and is antagonized by *HOX* products²⁶. This indicates

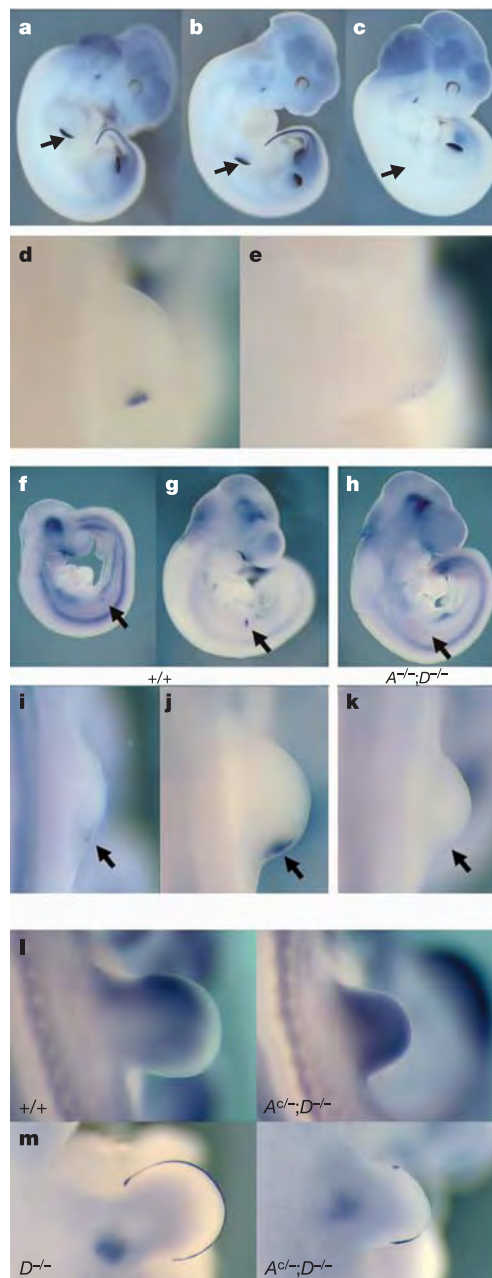


Figure 4 | Lack of *Shh* expression in double-mutant forelimbs. **a–c**, At day 11, control animals (**a**) express *Shh* at the posterior margin of the bud (arrow). Embryos lacking *HoxD* and one copy of *HoxA* ($A^{+/-};D^{-/-}$) (**b**) still have a normal *Shh*. However, the conditional double mutant ($A^{c/-};D^{-/-}$) no longer shows *Shh* transcripts in forelimbs (**c**), unlike in hindlimbs where the deletion is incomplete. **d, e**, *Hox* genes maintain *Shh* transcription. At E10, some *Shh* expression is observed in conditional double-mutant ($A^{c/-};D^{-/-}$) forelimbs (**d**), as a result of residual *HoxA* transcripts produced before completion of the deletion. In contrast, E11 buds (**e**) contain only few cells, if any, expressing *Shh*. **f–k**, *Shh* is not expressed in forelimbs fully deficient for both the *HoxA* and *HoxD* clusters (**h, k**), despite the delayed growth, as in both aged-matched (**g, j**) and younger (**f, i**) wild-type limb buds *Shh* expression is unambiguously detected. **l, m**, Expression of *Gli3* (**l**) and *Fgf8* (**m**) in double conditional-mutant ($A^{c/-};D^{-/-}$) forelimb buds. *Gli3* is expressed throughout the bud, unlike in the control, and *Fgf8* expression in the AER is patchy and interrupted. In both cases, transcript patterns resemble those observed in mice lacking *Shh* function^{16,17}.

that a generic animal appendage might be divided into two pieces: an ancestral, 'trunk'-dependent extension characterized by both the expression of *Meis* (*hth*) and the absence of *Shh* (*Hh*), whereas more distal parts evolved independently in various taxa²⁷, following different strategies, but only through the repression of *Meis* gene activity²⁵. In tetrapods, the same *Hox/Shh* module was co-opted in both the forelimbs and the hindlimbs, which provides an explanation for serial homology and subsequent concerted evolution of paired appendages²⁸. We therefore propose that the basic proximal/distal distinction proposed for arthropods⁶ be extended to tetrapods. The position of this morphological transition, in both arthropods and vertebrates, might nevertheless not correspond precisely to an articulation. Instead, it might occur at a developmental stage at which these morphological landmarks are not yet formed, leading to the above uncertainty in identifying the remaining structures present in the *HoxA/HoxD* deficient forelimbs.

METHODS

To generate mice with their *HoxA* complex flanked with *loxP* sites, we used ES cells (a gift from F. Rijli) containing a *loxP-TKneomycin-loxP* cassette downstream of *Hoxa1* (ref. 29) and further electroporated with a *loxP-pGKhygromycin* cassette at an *HpaI* site 3.9 kilobases (kb) upstream of the *Hoxa13*. The resulting floxed allele (*HoxA^{fllox}*) was introduced into the germ line of chimaeric males. The conditional deletion in limbs was obtained with mice expressing the Cre recombinase under the control of the *Prx1* promoter¹⁰. All animals and embryos were genotyped by Southern blot analysis. The *HoxD^{Del9-12}* mice were produced with the TAMERE strategy (B.T. and D.D., unpublished work). Whole-mount *in situ* hybridizations, skeletal preparations and cartilage staining were performed with standard procedures and previously described riboprobes.

Received 27 January; accepted 18 April 2005.

- Zakany, J., Kmita, M. & Duboule, D. A dual role for Hox genes in limb anterior-posterior asymmetry. *Science* **304**, 1669–1672 (2004).
- Davis, A. P., Witte, D. P., Hsieh-Li, H. M., Potter, S. S. & Capecchi, M. R. Absence of radius and ulna in mice lacking *hoxa-11* and *hoxd-11*. *Nature* **375**, 791–795 (1995).
- Fromental-Ramain, C. *et al.* *Hoxa-13* and *Hoxd-13* play a crucial role in the patterning of the limb autopod. *Development* **122**, 2997–3011 (1996).
- Favier, B. *et al.* Functional cooperation between the non-paralogous genes *Hoxa-10* and *Hoxd-11* in the developing forelimb and axial skeleton. *Development* **122**, 449–460 (1996).
- Wellik, D. M. & Capecchi, M. R. *Hox10* and *Hox11* genes are required to globally pattern the mammalian skeleton. *Science* **301**, 363–367 (2003).
- Snodgrass, R. E. *Principles of Insect Morphology* (McGraw-Hill, New York, 1935).
- Riddle, R. D., Johnson, R. L., Laufer, E. & Tabin, C. Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* **75**, 1401–1416 (1993).
- Suemori, H. & Noguchi, S. Hox C cluster genes are dispensable for overall body plan of mouse embryonic development. *Dev. Biol.* **220**, 333–342 (2000).
- Medina-Martinez, O., Bradley, A. & Ramirez-Solis, R. A large targeted deletion of *Hoxb1-Hoxb9* produces a series of single-segment anterior homeotic transformations. *Dev. Biol.* **222**, 71–83 (2000).
- Logan, M. *et al.* Expression of Cre recombinase in the developing mouse limb bud driven by a *Prx1* enhancer. *Genesis* **33**, 77–80 (2002).
- Spitz, F. *et al.* Large scale transgenic and cluster deletion analysis of the *HoxD* complex separate an ancestral regulatory module from evolutionary innovations. *Genes Dev.* **15**, 2209–2214 (2001).
- Rallis, C. *et al.* *Tbx5* is required for forelimb bud formation and continued outgrowth. *Development* **130**, 2741–2751 (2003).
- van den Akker, E. *et al.* Axial skeletal patterning in mice lacking all paralogous group 8 Hox genes. *Development* **128**, 1911–1921 (2001).
- Fromental-Ramain, C. *et al.* Specific and redundant functions of the paralogous *Hoxa-9* and *Hoxd-9* genes in forelimb and axial skeleton patterning. *Development* **122**, 461–472 (1996).
- Harfe, B. D. *et al.* Evidence for an expansion-based temporal *Shh* gradient in specifying vertebrate digit identities. *Cell* **118**, 517–528 (2004).
- Chiang, C. *et al.* Manifestation of the limb prepatter: limb development in the absence of sonic hedgehog function. *Dev. Biol.* **236**, 421–435 (2001).
- Kraus, P., Fraidenaich, D. & Loomis, C. A. Some distal limb structures develop in mice lacking Sonic hedgehog signalling. *Mech. Dev.* **100**, 45–58 (2001).
- Lewis, P. M. *et al.* Cholesterol modification of sonic hedgehog is required for long-range signalling activity and effective modulation of signaling by Ptc1. *Cell* **105**, 599–612 (2001).
- Ros, M. A. *et al.* The chick oligozeugodactyly (*ozd*) mutant lacks sonic hedgehog function in the limb. *Development* **130**, 527–537 (2003).
- Knezevic, V. *et al.* *Hoxd-12* differentially affects preaxial and postaxial chondrogenic branches in the limb and regulates Sonic hedgehog in a positive feedback loop. *Development* **124**, 4523–4536 (1997).
- Litingtung, Y., Dahn, R. D., Li, Y., Fallon, J. F. & Chiang, C. *Shh* and *Gli3* are dispensable for limb skeleton formation but regulate digit number and identity. *Nature* **418**, 979–983 (2002).
- te Welscher, P. *et al.* Progression of vertebrate limb development through SHH-mediated counteraction of *GLI3*. *Science* **298**, 827–830 (2002).
- Sordino, P. D. D. A molecular approach to the evolution of vertebrate paired appendages. *Trends Ecol. Evol.* **11**, 114–119 (1996).
- Gonzalez-Crespo, S. *et al.* Antagonism between extradenticle function and Hedgehog signalling in the developing limb. *Nature* **394**, 196–200 (1998).
- Mercader, N. *et al.* Conserved regulation of proximodistal limb axis development by *Meis1/Hth*. *Nature* **402**, 425–429 (1999).
- Capdevila, J., Tsukui, T., Rodriguez Esteban, C., Zappavigna, V. & Izpisua Belmonte, J. C. Control of vertebrate limb outgrowth by the proximal factor *Meis2* and distal antagonism of BMPs by *Gremlin*. *Mol. Cell* **4**, 839–849 (1999).
- Shubin, N., Tabin, C. & Carroll, S. Fossils, genes and the evolution of animal limbs. *Nature* **388**, 639–648 (1997).
- Tabin, C. J. & Laufer, E. Hox genes and the evolutionary origin of the serial homology between fore and hind limbs. *Nature* **361**, 692–693 (1993).
- Dupe, V. *et al.* *In vivo* functional analysis of the *Hoxa-1 3'* retinoic acid response element (3' RARE). *Development* **124**, 399–410 (1997).

Acknowledgements We thank A. McMahon, M. Torres, G. Martin, A. Ruiz I Altaba, C. Fromental-Ramain and E. Olson for probes, N. Fraudeau and M. Friedli for technical assistance, and J. Deschamps, G. Morata and P. Vassalli for discussions. This work was supported by funds from the canton de Genève, the Claraz and Louis Jeantet foundations, the Swiss National Research Fund, the NCCR 'Frontiers in Genetics' (to D.D.), a grant from the National Institutes of Health (to C.T.) and the EMBO young investigator program (to M.L.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.D. (denis.duboule@zoo.unige.ch).

Abnormal display of PfEMP-1 on erythrocytes carrying haemoglobin C may protect against malaria

Rick M. Fairhurst¹, Dror I. Baruch¹, Nathaniel J. Brittain¹, Graciela R. Ostera¹, John S. Wallach¹, Holly L. Hoang⁴, Karen Hayton¹, Aldiouma Guindo⁵, Morris O. Makobongo², Owen M. Schwartz³, Anatole Tounkara⁶, Ogobara K. Doumbo⁵, Dapa A. Diallo⁵, Hisashi Fujioka⁷, May Ho⁴ & Thomas E. Wellems¹

Haemoglobin C, which carries a glutamate-to-lysine mutation in the β -globin chain, protects West African children against *Plasmodium falciparum* malaria^{1,2}. Mechanisms of protection are not established for the heterozygous (haemoglobin AC) or homozygous (haemoglobin CC) states. Here we report a marked effect of haemoglobin C on the cell-surface properties of *P. falciparum*-infected erythrocytes involved in pathogenesis. Relative to parasite-infected normal erythrocytes (haemoglobin AA), parasitized AC and CC erythrocytes show reduced adhesion to endothelial monolayers expressing CD36 and intercellular adhesion molecule-1 (ICAM-1). They also show impaired rosetting interactions with non-parasitized erythrocytes, and reduced agglutination in the presence of pooled sera from malaria-immune adults. Abnormal cell-surface display of the main variable cytoadherence ligand, PfEMP-1 (*P. falciparum* erythrocyte membrane protein-1), correlates with these findings. The abnormalities in PfEMP-1 display are associated with markers of erythrocyte senescence, and are greater in CC than in AC erythrocytes. Haemoglobin C might protect against malaria by reducing PfEMP-1-mediated adherence of parasitized erythrocytes, thereby mitigating the effects of their sequestration in the microvasculature.

Malaria continues to kill more than a million African children annually. Over thousands of years, evolutionary pressure has selected a variety of haemoglobin mutations that confer resistance to severe manifestations of the disease³. The mechanisms by which these mutations exert their influence are largely unclear. Haemoglobin C (HbC) contains a glutamate-to-lysine mutation at the sixth position in the β -globin chain, the same position as the glutamate-to-valine mutation in sickle haemoglobin (HbS). HbC occurs mostly in West Africa, where its prevalence in some regions is more than 25% and greatly exceeds that of HbS⁴. Epidemiological studies have shown substantial malaria protection by both the AC and CC phenotypes in the Dogon and Mossi ethnic populations of Mali and Burkina Faso^{1,2}. Although reduced parasite proliferation in CC erythrocytes has been proposed as a mechanism of protection^{5,6}, *P. falciparum* grows normally in AC erythrocytes *in vitro*⁵. Furthermore, the frequent presence of substantial parasite densities in malarious AC and CC children indicates that a process of protection other than reduced proliferation must operate *in vivo*¹.

P. falciparum parasites remodel their host erythrocytes extensively, placing PfEMP-1 cytoadherence proteins in knob-like protrusions at the surface of the host cell membrane. These cell-surface modifications enable mature parasitized erythrocytes to sequester in the

microvasculature and avoid clearance from the bloodstream by the spleen. The adherence properties of parasitized erythrocytes can lead to life-threatening manifestations of disease—accumulation of parasitized erythrocytes in the brain, for example, produces the clinical syndrome of cerebral malaria⁷. Because of this link between erythrocyte-surface modifications and disease, we compared *P. falciparum*-infected AC, CC and AA erythrocytes in assays of cytoadherence, rosetting and agglutination, three phenomena that depend on PfEMP-1 expression at the host erythrocyte surface^{8,9}.

In the first of these comparisons, we tested the ability of parasitized erythrocytes to adhere to endothelial monolayers expressing two important host cytoadherence receptors, CD36 and ICAM-1. Under flow conditions¹⁰, parasitized AC erythrocytes showed ~25% reduced adherence to these endothelial monolayers compared with parasitized AA erythrocytes, and parasitized CC erythrocytes showed no adherence (Fig. 1a).

Rosette formation by attachment of non-parasitized to parasitized erythrocytes is likewise mediated by PfEMP-1 and involves host cell receptors, including complement receptor 1 (ref. 9). Increased rosette frequencies are associated with increased malaria severity¹¹. Using two different *P. falciparum* lines, we found that rosette frequencies were ~33% lower for AC erythrocytes compared with AA erythrocytes, and that rosettes formed only rarely with CC erythrocytes (Fig. 1b).

In our third set of comparisons, we tested unfixed parasitized erythrocytes for their ability to agglutinate in the presence of serum from individuals with immunity to malaria. This clumping in response to antibody–antigen interactions is a sensitive assay for the cell-surface expression of PfEMP-1 (refs 8, 12, 13). Immune sera agglutinated parasitized AC erythrocytes significantly less than parasitized AA erythrocytes, and agglutinates of parasitized CC erythrocytes were infrequent (Fig. 1c). Marked relative reductions occurred over a wide range of serum dilutions, confirming that the results had not been compromised by non-optimal ratios of antibodies to parasite antigens (Fig. 1d). These data are consistent with reduced antibody bridging of PfEMP-1 molecules between parasitized HbC erythrocytes.

To test whether HbC affects PfEMP-1-mediated phenomena *in vivo*, we studied parasitized AC and AA erythrocytes obtained directly from Malian children with malaria. After placing these samples in culture for 24 h to allow for the development of late-stage parasites expressing PfEMP-1, we tested for agglutination using sera pooled from nine immune adults in the same village. The parasitized erythrocytes from 19 out of 20 AA children agglutinated,

¹Laboratory of Malaria and Vector Research, ²Malaria Vaccine Development Branch, and ³Biological Imaging Facility, Research Technologies Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. ⁴Department of Microbiology and Infectious Diseases, University of Calgary, Calgary, Alberta T2N 4N1, Canada. ⁵Malaria Research and Training Center, Departments of Parasitology and Hematology, Faculty of Medicine, Pharmacy, and Odontostomatology, University of Bamako, Bamako BP 1805, Mali. ⁶Centre National de Transfusion Sanguine, Ministry of Health, Bamako BP E344, Mali. ⁷Institute of Pathology, Case Western Reserve University, Cleveland, Ohio 44106, USA.

confirming broad reactivity of the adult serum pool (Fig. 1e). In contrast, three out of three infected AC samples did not show sufficient numbers of agglutinates to yield a positive agglutination score. The mean agglutination scores of parasitized AC and AA erythrocytes were significantly different (0.0 versus 1.6; $P = 0.005$). In parallel experiments, we purified *P. falciparum* trophozoites from AA samples that had yielded agglutination scores of 4 and then inoculated the recovered parasites into fresh AC erythrocytes (two donors) or CC erythrocytes (three donors). After a single two-day cycle of reinvasion and growth, the resulting trophozoite-infected AC or CC erythrocytes gave agglutination scores of zero, providing additional evidence for an important effect of HbC on the antibody–PfEMP-1 bridges between parasitized erythrocytes. This effect appeared to involve PfEMP-1 display specifically, as control experiments showed no differences between the agglutination of AA, AC, and CC erythrocytes using standard ABO/Rh blood typing antibodies.

Altered adherence properties of HbC erythrocytes relative to AA erythrocytes could be due to reduced levels of PfEMP-1 or to its altered distribution in knobs, both of which can exert important influences on the organization and cooperativity of PfEMP-1 interactions with attaching molecules¹⁴. To explore these possibilities, we used flow cytometry to examine unfixed parasitized erythrocytes for their reactivity to rat antisera against different forms of PfEMP-1 expressed by the mature trophozoite stages of two *P. falciparum* lines, MC/R+ and FVO. Forty-eight hours after infection with purified, synchronized trophozoites, populations of AA and AC erythrocytes showed overlapping distributions of fluorescence, with parasitized AC erythrocytes showing ~15% lower median fluorescence intensities (MFI) (Fig. 2 and Supplementary Table 1; mean AC:AA MFI ratio $\pm$ s.d. = 0.85 ± 0.10 ; $P < 0.0001$, two-tailed, one sample *t*-test of the mean). MFIs from parasitized CC erythrocytes were markedly lower (Fig. 2c, f and Supplementary Table 1; mean CC:AA MFI ratio $\pm$ s.d. = 0.33 ± 0.12 ; $P < 0.0001$). Notably, the MC/R+ and FVO *P. falciparum* lines consistently showed significant

differences in their expression of PfEMP-1 on CC erythrocytes (Fig. 2c, f; 8% of parasitized erythrocytes expressing PfEMP-1 for MC/R+ compared with 38% for FVO). These findings might reflect differing abilities of *P. falciparum* lines to develop in and remodel the surface of CC erythrocytes *in vitro*⁶, and raise the possibility that protection against malaria by HbC is strain-dependent.

In parallel with flow cytometry, we used confocal microscopy to examine the distribution of PfEMP-1 on the surface of fresh AA, AC and CC erythrocytes infected with comparably mature *P. falciparum* trophozoites. Deconvoluted confocal images and cell-surface stack projections showed uniform and finely distributed signals from PfEMP-1 over the surfaces of AA erythrocytes (Fig. 3a). Parasitized CC erythrocytes yielded fluorescence intensities that were often weaker and showed irregular and widely dispersed PfEMP-1 distributions (Fig. 3b). On AC erythrocytes, PfEMP-1 showed a spectrum of intensities and distributions that in some cases were AA-like (Fig. 3c) and in other cases were CC-like (Fig. 3d). The presence of significant numbers of parasitized AC erythrocytes with CC-like PfEMP-1 display (~25% of the parasitized cells in our samples) is the probable explanation for the reduced MFI levels of AC populations measured by flow cytometry.

PfEMP-1 molecules are anchored in knobs placed by *P. falciparum*

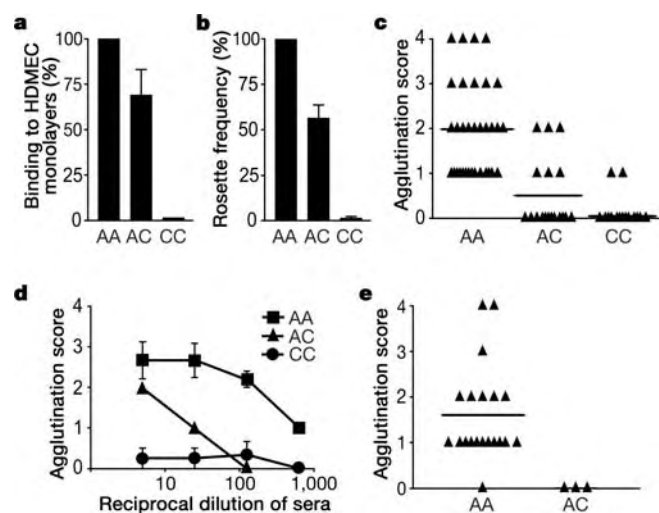


Figure 1 | Cytoadherence, rosetting and agglutination of *P. falciparum*-infected erythrocytes. **a**, Adherence of parasitized AA, AC, and CC erythrocytes to human dermal microvascular endothelial cell (HDMEC) monolayers under flow conditions. **b**, Rosette frequencies of erythrocyte samples (4 AA, 3 AC, 1 CC) parasitized with *P. falciparum* lines (MC/R+ or TM284). **c**, Agglutination scores from erythrocyte samples (5 AA, 2 AC, 3 CC) parasitized with *P. falciparum* lines (7G8, GB4, FCR3 or MC/R+) and tested with eight antisera (not all combinations used). **d**, Agglutination scores of erythrocyte samples (5 AA, 2 AC, 3 CC) parasitized with *P. falciparum* lines (7G8 or GB4) and tested with three antisera at serial dilutions. **e**, Agglutination scores of naturally infected AA and AC erythrocytes from Malian children. Error bars show s.d.

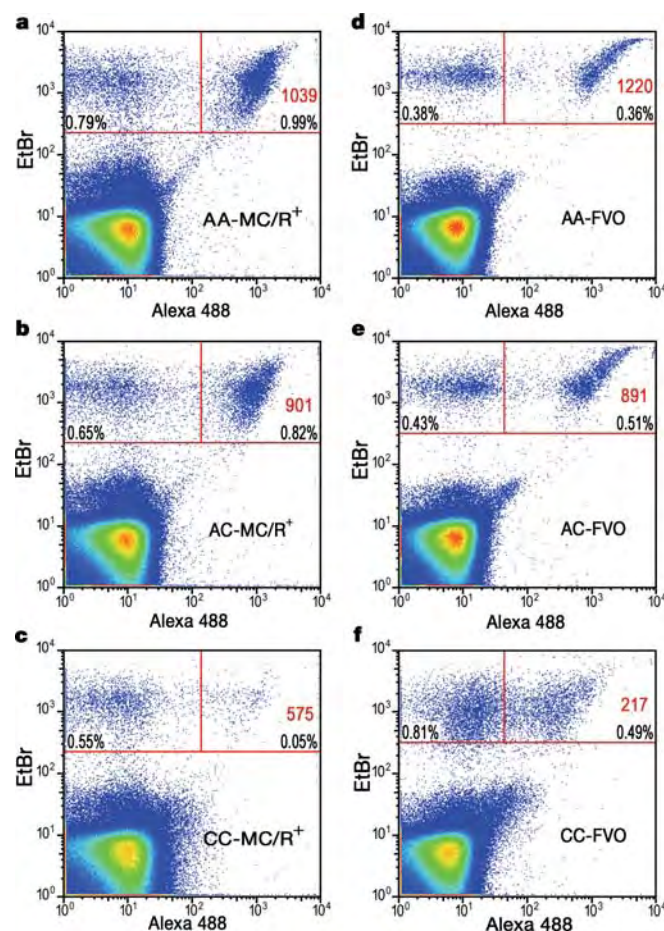


Figure 2 | Flow cytometry analysis of PfEMP-1 expression on normal and HbC erythrocytes parasitized with *P. falciparum*. **a–f**, Parasitized erythrocytes were stained with ethidium bromide (EtBr) and probed with a rat polyclonal antiserum specific for an extracellular PfEMP-1 molecule of *P. falciparum* line MC/R+ (**a–c**) or FVO (**d–f**). Parasitized erythrocytes reactive to antisera appear as clustered populations in the upper right quadrants (red numbers indicate MFI). Upper left quadrants include signals from parasite populations that express variant PfEMP-1 molecules not reactive with antisera. Percentage figures in the upper quadrants were calculated from the total erythrocyte number in each sample.

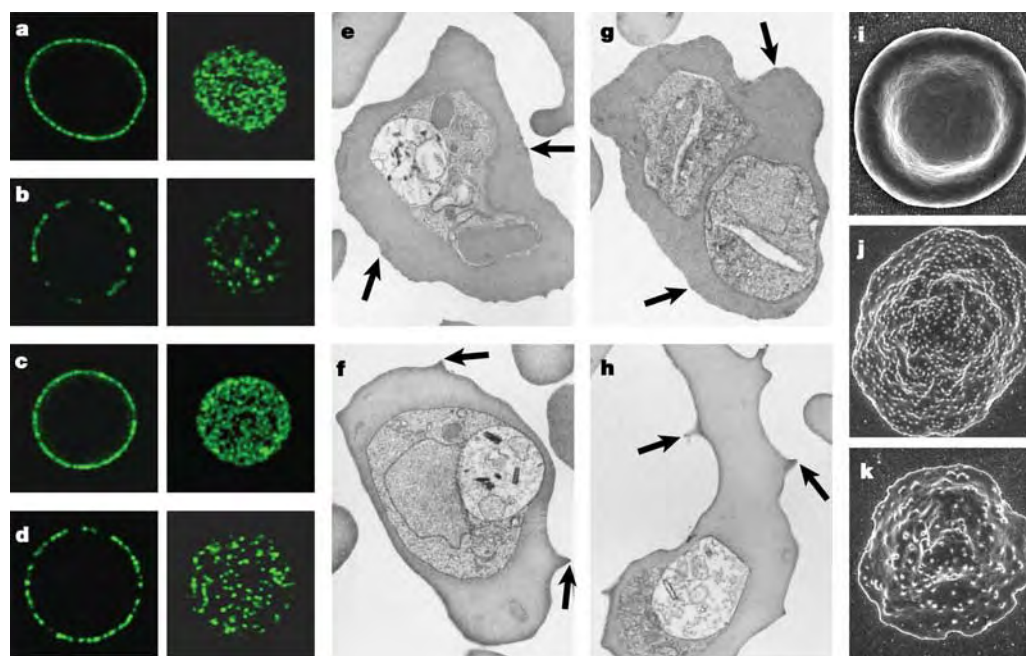


Figure 3 | PfEMP-1 and knob distributions on normal and HbC erythrocytes parasitized with *P. falciparum*. **a–d**, Confocal cross-sections through the midplane (left column) or maximum intensity projections of z stacks through the upper cell surface (right column) of trophozoite-infected erythrocytes probed with a polyclonal antiserum against PfEMP-1 (FVO line). **a**, PfEMP-1 fluorescence patterns typical of a parasitized AA erythrocyte. **b**, Irregular, patchy PfEMP-1 distributions typical of a parasitized CC erythrocyte. **c**, Fluorescence patterns from a parasitized AC erythrocyte, similar to those of parasitized AA erythrocytes. **d**, Dispersed and brightly punctate fluorescence signals from abnormally displayed

PfEMP-1 on an AC erythrocyte. **e, f**, Transmission electron micrographs of parasitized AC erythrocytes in continuous culture. Arrows indicate knobs that have a normal (AA-like) appearance (**e**) or abnormal (CC-like⁶) appearance (**f**). Some parasitized AC erythrocytes were also observed to have both normal and abnormal knobs (data not shown). **g, h**, Knobs with normal (**g**) or abnormal (**h**) appearance on a parasitized AC erythrocytes from a child with malaria. **i–k**, Scanning electron micrographs showing a normal uninfected AA erythrocyte (**i**), frequent knobs on a trophozoite-infected AA erythrocyte (**j**), and abnormally large and widely separated knobs on a similarly parasitized AC erythrocyte (**k**).

under the host erythrocyte membrane¹⁵, and their display at the cell surface is affected when knob placement and organization is abnormal¹⁴. We therefore evaluated the ultrastructural appearance of knobs on trophozoite-infected AA and AC erythrocytes using both transmission and scanning electron microscopy. In transmission electron micrographs, we observed distinct populations of parasitized AC erythrocytes: (1) a population with regularly distributed arrays of fine knobs resembling those of parasitized AA erythrocytes (Fig. 3e) and (2) a population with abnormally large, widely separated knobs that resembled those of parasitized CC erythrocytes⁶ (~25% of the infected AC cells; Fig. 3f) and were not observed in control parasitized AA erythrocytes. We also examined parasitized erythrocytes drawn from malaria patients and cultivated within 30 h to mature trophozoite stages (Fig. 3g, h). These again showed that ~25% of parasitized AC but not AA erythrocytes carried large and widely separated knobs (Fig. 3h). Scanning electron microscopy confirmed extensive remodelling of the surface of parasitized AA and AC erythrocytes relative to the smooth, biconcave surface of uninfected erythrocytes (Fig. 3i). The knobs on parasitized AA erythrocytes were numerous and finely distributed (Fig. 3j), whereas parasitized AC erythrocytes frequently had larger and widely separated knobs (Fig. 3k). Concordant appearance of these images with the patterns of immunofluorescence signals (Fig. 3a, d) is consistent with widely dispersed anchoring and irregular concentrations of PfEMP-1 in the knobs of parasitized AC and CC erythrocytes. Direct studies of PfEMP-1 distributions in relation to the knob topography of parasitized AC and CC erythrocytes, including examination by atomic force microscopy¹⁶, will be required to confirm and extend these observations.

The frequent presence of abnormal knobs on AC erythrocytes might be associated with the cellular changes of erythrocyte ageing¹⁷. Such age-associated effects include increases in cell density and mean

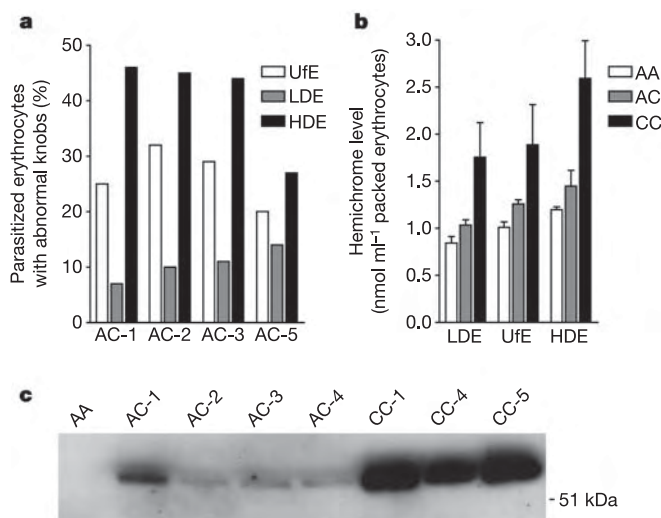


Figure 4 | Effects of haemoglobin C on knob formation, hemichrome deposition and senescent antibody binding *in vivo*. **a**, Prevalence of abnormal knobs in low- and high-density erythrocytes from four different AC samples parasitized with *P. falciparum* clone 7G8. UfE, unfractionated erythrocytes; LDE, low-density erythrocytes; HDE, high-density erythrocytes. **b**, Hemichrome levels in low- or high-density fractions of erythrocytes from 3 AA, 3 AC and 3 CC donors. Error bars indicate s.d. **c**, Senescent antibody signals from freshly drawn erythrocytes of one AA, 4 AC and 3 CC donors. Three additional AA samples also gave no signal (data not shown).

corpuscular haemoglobin concentration (MCHC). AC erythrocyte populations contain 30–40% HbC and have MCHCs ($34.2 \pm 1.7 \text{ g dl}^{-1}$) midway between the concentration in normal AA erythrocytes ($31.8 \pm 1.2 \text{ g dl}^{-1}$) and the more dense CC erythrocytes ($36.4 \pm 1.4 \text{ g dl}^{-1}$) (refs 18–20). The range of densities about these mean values allows separation of erythrocyte populations into high- and low-density fractions by ultracentrifugation in autologous plasma²¹. When we performed these separations and used transmission electron microscopy to examine the knobs of AC erythrocytes of different densities infected with mature trophozoites, we found that the proportion of parasitized erythrocytes displaying abnormal knobs was an average of 4 times greater in high-density than in low-density AC fractions (Fig. 4a; mean $\pm$ s.d. $40.5 \pm 9.0\%$ for high-density versus $10.5 \pm 2.9\%$ for low-density AC erythrocytes; $P = 0.02$).

Erythrocyte senescence is associated with the oxidation of haemoglobin to hemichromes, which lodge beneath the membrane of the cell and decrease its plasticity. Clustering of band 3 by hemichromes induces binding of autologous immunoglobulin, fixation of complement and opsonization of senescent erythrocytes²². We examined fresh AC and CC erythrocytes for elevated levels of hemichromes and membrane-bound autoantibody. Results showed increased levels of hemichromes in CC and high-density AC erythrocytes relative to AA and low-density AC erythrocytes (Fig. 4b). Surface autoantibody was detected on the erythrocytes of all AC donors and, at much higher levels, on those of CC donors; in contrast, AA erythrocytes carried no detectable autoantibody (Fig. 4c). An increased susceptibility of HbC to oxidation²³ might underlie these and other cellular changes that compromise the ability of *P. falciparum* to remodel HbC erythrocytes.

PfEMP-1 has a central role in the cytoadherence and rosetting of parasitized erythrocytes and the effects of these phenomena in severe malaria^{24,25}. We suggest that haemoglobin C affects the display of PfEMP-1 by reducing both its expression level and distribution at the host cell surface. These effects would in turn mitigate obstruction and inflammation resulting from the adherence of parasitized erythrocytes in the microvasculature, and might be complemented by an increased rigidity of HbC erythrocytes. This model of malaria protection by HbC has features independent of acquired immunity, and could therefore be implicated in the protection of young children and pregnant women, who often lack the variant-specific PfEMP-1 antibodies associated with clinical protection^{12,13,26}. Varying spectra of PfEMP-1 abnormalities on parasitized AC and CC erythrocytes can help to explain why some children with HbC still suffer severe malaria and why the CC phenotype might afford greater protection than the AC phenotype, as some results have suggested². As with other malaria-protective polymorphisms, such as those of sickle trait and glucose-6-phosphate dehydrogenase (G6PD) deficiency²⁷, HbC protection can vary among human populations¹. Effects of genetic background, parasite strain variations, regional patterns of disease transmission and incidence, nutritional status and acquired immunity all might influence the susceptibility of HbC carriers to severe malaria. Identification of factors that modify the AC and CC phenotypes will improve our understanding of malaria-protective mechanisms and help to explain their variable fitnesses across the mutant haemoglobin clines of Africa and southeast Asia⁴.

METHODS

Red blood cells. Heparinized blood was washed with RPMI 1640 (Invitrogen/Gibco) and stored at 50% hematocrit at 4 °C before use. Haemoglobin genotypes were determined by cellulose acetate and citrate agar electrophoresis, and confirmed by high-performance liquid chromatography. Blood collection was approved by the Institutional Review Boards of the National Institute of Allergy and Infectious Diseases and the University of Bamako.

Parasite culture. Parasites in AA and AC erythrocytes were cultivated at 5% hematocrit in complete medium (RPMI 1640 supplemented with 25 mg ml⁻¹ HEPES, 2 mg ml⁻¹ sodium bicarbonate, 50 µg ml⁻¹ gentamicin and 10% (v/v)

heat-inactivated human O⁺ or AB⁺ serum). Parasites in CC erythrocytes were cultivated similarly but at lower (0.5–1.0%) hematocrit. This modification allowed us to maintain parasite cultures in CC erythrocytes that were synchronous with control cultures in AA erythrocytes (see Supplementary Table 2) and contained >90% viable forms, as determined by transmission electron microscopy. All cultures were maintained at 37 °C in an atmosphere of 5% CO₂, 5% O₂ and 90% N₂, with twice daily medium changes. Erythrocytes were obtained in both the United States and Mali. In all experiments, age-matched AA, AC and CC erythrocytes were simultaneously inoculated by the same parasite stock within 4–48 h of blood collection, as described⁶. Parasites expressing knobs were maintained by periodic gelatin flotation⁶. *P. falciparum*-infected erythrocytes from Malian children were placed in culture for 24 h to allow development of late-trophozoite stage parasites suitable for agglutination assays (see below).

Flow cytoadherence assays. Human dermal microvascular endothelial cells (HDMEC) were harvested from discarded neonatal human foreskin samples following a protocol approved by The Conjoint Ethics Review Board of the University of Calgary. HDMEC were maintained in endothelial basal medium and grown to confluent monolayers as described¹⁰. Suspensions of parasitized erythrocytes (1% hematocrit, 5% parasitemia in RPMI 1640, pH 7.2) were drawn through a parallel-plate flow chamber at a rate producing shear stresses approximating those in the microvasculature¹⁰. After 7 min video recording, five additional fields were recorded (30 s each). Parasitized erythrocytes were considered adherent only if they remained stationary for more than 10 s. Mean counts were obtained from at least five microscopic fields and were expressed as the number of adherent parasitized erythrocytes per mm² surface area. The numbers of infected AC and CC erythrocytes bound were calculated as percentages of the number of control infected AA erythrocytes bound. Experimental and control erythrocytes were tested simultaneously. Erythrocytes from two AA and two AC donors were infected with the *P. falciparum* 7G8 line and tested for binding on the same cell passage of two different HDMEC monolayer preparations. 7G8-parasitized CC erythrocytes were tested on three sequential passages of the same HDMEC preparation.

Rosetting. MC/R⁺ and TM284 parasites were cultured to 5–10% parasitemia in AA, AC and CC erythrocytes, as above. Erythrocytes containing mature parasites were identified by their refractile hemazoin pigment. Rosettes were identified as parasitized erythrocytes to which three or more uninfected erythrocytes were closely adhered. Absolute rosette rates were determined by dividing the total number of rosettes by the number of parasitized erythrocytes examined (>200). Rosette frequencies (%) were calculated relative to the rosette rates of control parasitized AA erythrocytes.

Parasitized erythrocyte agglutination assays. Agglutination of parasitized erythrocytes (5% at mature-stage parasitemia) was tested and scored as previously described²⁸. Immune sera included pooled adult serum samples from various regions of Africa endemic for *P. falciparum* malaria, and convalescent serum from *Pan troglodytes* experimentally infected with a genetic cross of the *P. falciparum* lines GB4 and 7G8 (unpublished). Use of relevant non- and pre-immune sera resulted in no agglutination.

Flow cytometry. Rat polyclonal antisera were raised against externalized segments of PfEMP-1 expressed by the MC/R⁺ and FVO parasite lines. Approximately 1×10^6 erythrocytes at 1% parasitemia were stained with antisera (anti-MC/R⁺ at 1:80 dilution; anti-FVO at 1:800) in PBS containing 2% fetal calf serum for 30 min at 25 °C and washed. Bound antibody was detected using Alexa 488-conjugated anti-rat IgG (Molecular Probes). Parasitized erythrocytes were stained with ethidium bromide (2 µg ml⁻¹) for 30 min. A FACSort instrument (Becton-Dickinson) and FlowJo software were used to acquire and analyse 500,000 events from each sample.

Surface immunofluorescence/confocal microscopy. Parasitized erythrocytes (2 µl packed cell volume) were added to 20 µl polyclonal antiserum (rat, 1:100 dilution, MC/R⁺; 1:1,500 dilution, FVO) for 1 h at 25 °C and washed. Bound antibody was detected with Alexa 488-conjugated anti-rat IgG. Control experiments using antiserum against the intracellular acidic terminal segment (ATS) of PfEMP-1 gave no signal, confirming that the unfixed, parasitized erythrocytes in these experiments were not antibody permeable. Images were collected using a TCS-SP2 AOBs confocal microscope (Leica Microsystems) using a $\times 63$ oil immersion objective NA 1.4, zoom 6. The confocal pinhole was set to 0.9 Airy units to ensure maximum resolution. Alexa 488 fluorescence was excited using an argon laser at 488 nm. DIC (differential interference contrast) images were collected using the same 488-nm excitation wavelength but with the transmitted light detector. 3-dimensional reconstructions were made using sequential sections through the sample with a z increment of 0.12 µm. Images were deconvoluted and processed using Leica TCS (version 2.1374), Imaris 4.1 (3-D reconstructions) (Bitplane AG), Huygens Essentials (deconvolution) (SVI) and Adobe Photoshop CS (Adobe Systems).

Electron microscopy. Parasites were processed for transmission electron microscopy as described⁶. The knobs of 50–200 trophozoite-infected erythrocytes were scored from each sample. For scanning electron microscopy, *P. falciparum*-infected AA and AC erythrocytes were washed in phosphate-buffered saline (pH 7.4), fixed in 2.5% glutaraldehyde plus 4% paraformaldehyde in 0.1 M sodium cacodylate/0.05 M sucrose. Samples were photographed at 5 kV using a Hitachi S-4500 field emission scanning electron microscope.

Density fractionation of AA and AC erythrocytes. Erythrocyte suspensions were adjusted to 90% hematocrit in autologous plasma and ultracentrifuged at 51,000g for 2 h at 25 °C (ref. 21). The top and bottom 10% or 25% volumes of packed erythrocytes were removed and used in parasite cultures or hemichrome extractions.

Quantification of hemichromes and detection of senescent antibody. Heparinized blood was washed in HEPES buffer, pH 7.4 (10 mM HEPES, 140 mM NaCl, 10 mM glucose) and the erythrocytes were lysed with Tris-HCl/EDTA (pH 8.0) in the presence of protease inhibitors (Complete; Roche Diagnostics GmbH). Membrane pellets were extracted as described²⁹. Hemichromes were quantified using the Winterbourne equation³⁰. For detection of senescent antibody, protein extract (1 µg) was resolved by SDS-PAGE, transferred to a PVDF membrane, and probed with a horseradish peroxidase-conjugated anti-human IgG antibody and chemiluminescent substrate.

Received 31 August 2004; accepted 11 April 2005.

- Agarwal, A. *et al.* Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood* **96**, 2358–2363 (2000).
- Modiano, D. *et al.* Haemoglobin C protects against clinical *Plasmodium falciparum* malaria. *Nature* **414**, 305–308 (2001).
- Flint, J., Harding, R. M., Boyce, A. J. & Clegg, J. B. The population genetics of the haemoglobinopathies. *Baillieres Clin. Haematol.* **11**, 1–51 (1998).
- Allison, A. C. Genetic factors in resistance to malaria. *Ann. NY Acad. Sci.* **91**, 710–729 (1961).
- Friedman, M. J., Roth, E. F., Nagel, R. L. & Trager, W. The role of hemoglobins C, S, and Nbal in the inhibition of malaria parasite development *in vitro*. *Am. J. Trop. Med. Hyg.* **28**, 777–780 (1979).
- Fairhurst, R. M., Fujioka, H., Hayton, K., Collins, K. F. & Wellem, T. E. Aberrant development of *Plasmodium falciparum* in hemoglobin CC red cells: implications for the malaria protective effect of the homozygous state. *Blood* **101**, 3309–3315 (2003).
- Miller, L. H., Baruch, D. I., Marsh, K. & Doumbo, O. K. The pathogenic basis of malaria. *Nature* **415**, 673–679 (2002).
- Baruch, D. I. *et al.* Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82**, 77–87 (1995).
- Rowe, J. A., Moulds, J. M., Newbold, C. I. & Miller, L. H. *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. *Nature* **388**, 292–295 (1997).
- Yipp, B. G. *et al.* Synergism of multiple adhesion molecules in mediating cytoadherence of *Plasmodium falciparum*-infected erythrocytes to microvascular endothelial cells under flow. *Blood* **96**, 2292–2298 (2000).
- Rowe, A., Obeiro, J., Newbold, C. I. & Marsh, K. *Plasmodium falciparum* rosetting is associated with malaria severity in Kenya. *Infect. Immun.* **63**, 2323–2326 (1995).
- Marsh, K., Otoo, L., Hayes, R. J., Carson, D. C. & Greenwood, B. M. Antibodies to blood stage antigens of *Plasmodium falciparum* in rural Gambians and their relation to protection against infection. *Trans. R. Soc. Trop. Med. Hyg.* **83**, 293–303 (1989).
- Bull, P. C. *et al.* Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature Med.* **4**, 358–360 (1998).
- Crabb, B. S. *et al.* Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. *Cell* **89**, 287–296 (1997).
- Arie, S. S. *et al.* *Plasmodium falciparum* erythrocyte membrane protein 1 is anchored to the actin-spectrin junction and knob-associated histidine-rich protein in the erythrocyte skeleton. *Mol. Biochem. Parasitol.* **108**, 237–247 (2000).
- Arie, T., Fairhurst, R. M., Brittain, N. J., Wellem, T. E. & Dvorak, J. A. Hemoglobin C modulates the surface topology of *Plasmodium falciparum*-infected erythrocytes. *J. Struct. Biol.* **150**, 163–169 (2005).
- Williams, A. R. & Morris, D. R. The internal viscosity of the human erythrocyte may determine its lifespan *in vivo*. *Scand. J. Haematol.* **24**, 57–62 (1980).
- Huisman, T. H. Chromatographic separation of hemoglobins A 2 and C. The quantities of hemoglobin A 2 in patients with AC trait, CC disease, and C-β-thalassemia. *Clin. Chim. Acta* **40**, 159–163 (1972).
- Bunn, H. F. *et al.* Molecular and cellular pathogenesis of hemoglobin SC disease. *Proc. Natl Acad. Sci. USA* **79**, 7527–7531 (1982).
- Charache, S., Conley, C. L., Waugh, D. F., Ugoretz, R. J. & Spurrell, J. R. Pathogenesis of hemolytic anemia in homozygous hemoglobin C disease. *J. Clin. Invest.* **46**, 1795–1811 (1967).
- Murphy, J. R. Influence of temperature and method of centrifugation on the separation of erythrocytes. *J. Lab. Clin. Med.* **82**, 334–341 (1973).
- Low, P. S., Waugh, S. M., Zinke, K. & Drenckhahn, D. The role of hemoglobin denaturation and band 3 clustering in red blood cell aging. *Science* **227**, 531–533 (1985).
- MacDonald, V. W. & Charache, S. Drug-induced oxidation and precipitation of hemoglobins A, S and C. *Biochim. Biophys. Acta* **701**, 39–44 (1982).
- Kaul, D. K., Roth, E. F. Jr, Nagel, R. L., Howard, R. J. & Handunnetti, S. M. Rosetting of *Plasmodium falciparum*-infected red blood cells with uninfected red blood cells enhances microvascular obstruction under flow conditions. *Blood* **78**, 812–819 (1991).
- MacPherson, G. G., Warrell, M. J., White, N. J., Looareesuwan, S. & Warrell, D. A. Human cerebral malaria. A quantitative ultrastructural analysis of parasitized erythrocyte sequestration. *Am. J. Pathol.* **119**, 385–401 (1985).
- Ricke, C. H. *et al.* Plasma antibodies from malaria-exposed pregnant women recognize variant surface antigens on *Plasmodium falciparum*-infected erythrocytes in a parity-dependent manner and block parasite adhesion to chondroitin sulfate A. *J. Immunol.* **165**, 3309–3316 (2000).
- Ruwende, C. *et al.* Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* **376**, 246–249 (1995).
- Gamain, B. *et al.* The surface variant antigens of *Plasmodium falciparum* contain cross-reactive epitopes. *Proc. Natl Acad. Sci. USA* **98**, 2664–2669 (2001).
- Giribaldi, G., Ulliers, D., Mannu, F., Aresé, P. & Turrini, F. Growth of *Plasmodium falciparum* induces stage-dependent haemichrome formation, oxidative aggregation of band 3, membrane deposition of complement and antibodies, and phagocytosis of parasitized erythrocytes. *Br. J. Haematol.* **113**, 492–499 (2001).
- Winterbourne, C. C. in *CRC Handbook of Methods of Oxygen Radical Research* (ed. Greenwald, R. A.) 137–141 (CRC Press, Boca Raton, 1985).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank our blood donors and J. F. Casella, B. Link, G. Rodgers, M. Law, K. D. Luc, N. Murray, J. A. Dvorak, F. Tokumasu, T. Arie, E. Fischer, L. H. Miller, A. Sadou, A. Katilé, B. Dakouo, K. Traoré and S. A. S. Diakité for their efforts in support of this work. H. F. acknowledges support from the United States Agency for International Development and the National Institutes of Health.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.E.W. (twelless@niaid.nih.gov).

LETTERS

R gene expression induced by a type-III effector triggers disease resistance in rice

Keyu Gu^{1*}, Bing Yang^{2*}, Dongsheng Tian^{1*}, Lifang Wu^{1*}, Dongjiang Wang^{1†}, Chellamma Sreekala^{1†}, Fan Yang^{1†}, Zhaoqing Chu^{1†}, Guo-Liang Wang^{3†}, Frank F. White² & Zhongchao Yin¹

Disease resistance (*R*) genes in plants encode products that specifically recognise incompatible pathogens and trigger a cascade of events leading to disease resistance in the host plant¹. *R*-gene specificity is dictated by both host *R* genes and cognate avirulence (*avr*) genes in pathogens^{2,3}. However, the basis of gene-for-gene specificity is not well understood. Here, we report the cloning of the *R* gene *Xa27* from rice and the cognate *avr* gene *avrXa27* from *Xanthomonas oryzae* pv. *oryzae*. Resistant and susceptible alleles of *Xa27* encode identical proteins. However, expression of only the resistant allele occurs when a rice plant is challenged by bacteria harbouring *avrXa27*, whose product is a nuclear localized type-III effector. Induction of *Xa27* occurs only in the immediate vicinity of infected tissue, whereas ectopic expression of *Xa27* resulted in resistance to otherwise compatible strains of the pathogen. Thus *Xa27* specificity towards incompatible pathogens involves the differential expression of the *R* gene in the presence of the *AvrXa27* effector.

Many *R* genes corresponding to diverse pathogens have been isolated, and fall primarily into five classes³. The majority of *avr* genes identified from bacterial pathogens encode effectors, also known as Avr proteins, which are secreted by type-III secretion pathways and, in many cases, internalized in the host cells⁴. Bacterial blight, a disease affecting rice and caused by *X. oryzae* pv. *oryzae*, is a significant agronomic problem in many rice-growing regions and is an ideal model system for the study of the interaction between plants and their bacterial pathogens. Nearly thirty rice loci have been recognized for resistance to bacterial blight, and four have been cloned^{5–8}. However, no bacterial *avr* gene corresponding to the cloned rice *R* gene has been identified.

The *Xa27* locus of rice confers resistance to diverse strains of *X. oryzae* pv. *oryzae*, including PXO99^A (ref. 9). We cloned *avrXa27*, an *avr* gene corresponding to *Xa27*, from PXO99^A by transferring the gene to compatible recipients. The gene with *Xa27*-dependent elicitor activity resided in a region of the PXO99^A genome containing at least five similar genes, which differed according to the number of direct repeats in the central repetitive region (Fig. 1a). The genes are bounded by sequences related to phage and transposon genes, indicating that the genes were subjected to horizontal transfer and gene amplification. *avrXa27* encodes a protein of 1,136 amino-acid residues and 16.5 thirty-four amino-acid direct repeats (Fig. 1b and Supplementary Fig. 1a). *AvrXa27* is similar to members of the AvrBs3/PthA family of type-III effectors (not shown), and has a conserved carboxy-terminal region containing three nuclear localization signal (NLS) motifs and a transcription activation domain,

which are found in all members of the family (Fig. 1b). *avrXa27* inhibited bacterial growth on rice plants containing *Xa27* when introduced into the compatible field strain AXO1947 (Fig. 1c).

AvrXa27 is distinct from other AvrBs3/PthA family members based on the repeat organization (Supplementary Fig. 1b). *AvrBs3*, *AvrXa7* and *AvrXa10* differ primarily in the number and arrangement of the repeats, which contribute to specificity, and require nuclear localization and the activation domain for function^{10–13}. We replaced the repetitive region of *AvrXa10* with that of *AvrXa27*, resulting in a fully active effector with *Xa27*-dependent activity and no *AvrXa10* activity (not shown). Thus, the repetitive region of *AvrXa27*, at least in part, contributes to the specificity of the

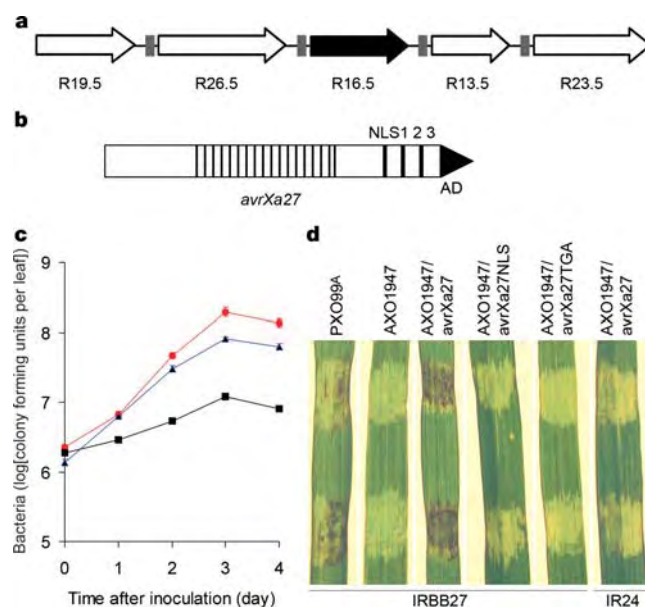


Figure 1 | *AvrXa27* is an AvrBs3/PthA type-III effector. **a**, *avrXa27* (black arrow) resides among four additional homologues (R indicates the number of repeats). Grey blocks indicate sequences related to a phage gene (NP_536678). **b**, *avrXa27* contains 16.5 repeats (thin boxes), three NLS motifs and an activation domain. **c**, Growth of AXO1947 with (black squares) and without *avrXa27* (red circles) in IRBB27, and with *avrXa27* in IR24 (blue triangles). **d**, *avrXa27* requires the NLS and activation-domain features for *Xa27*-dependent resistance. Resistance is indicated by tissue browning at the inoculation site within 48–72 h.

¹Temasek Life Sciences Laboratory, National University of Singapore, 1 Research Link, Singapore 117604, Singapore. ²Department of Plant Pathology, Kansas State University, Manhattan, Kansas 66506, USA. ³Institute of Molecular Agrobiotechnology, National University of Singapore, 1 Research Link, Singapore 117604, Singapore. [†]Present addresses: Laboratory of Plant Molecular Signalling, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China (D.W.); Department of Plant Physiology, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki, 305 8602, Japan (C.S.); Institute of Molecular and Cell Biology, 61, Biopolis Drive (Proteos), Singapore 138673, Singapore (F.Y.); Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore (Z.C.); and Department of Plant Pathology, Ohio State University, Columbus, Ohio 43210, USA (G.-L.W.).

*These authors contributed equally to this work.

Xa27-dependent resistance. To determine the NLS and activation-domain requirements, the respective coding sequences of *avrXa27* were replaced with non-consensus or truncated sequences (Supplementary Fig. 1c). In the absence of the NLS motifs, *avrXa27* did not confer the ability to elicit *Xa27*-dependent resistance to bacterial blight (Fig. 1d). Similarly, removal of the activation domain of *avrXa27* also resulted in the loss of resistance-elicitor activity (Fig. 1d and Supplementary Fig. 1c). Thus, *AvrXa27* is typical of this family of effectors that require NLS and activation-domain features and for which no *R* gene has been cloned¹⁴.

Xa27 was isolated from rice variety IRBB27 by map-based cloning. Two of sixteen subclones from the *Xa27* region on chromosome 6, AA17.6 and NN9.9, resulted in plants with *Xa27*-dependent resistance (Supplementary Table 1 and Supplementary Fig. 2c). Further subcloning associated resistance with the 5.2-kilobase (kb) *NsiI*/*AvrII* (NA5.2), 7.0-kb *PstI* (PP7) and 3.2-kb *NsiI*/*ApaI* (NA3.2) fragments,

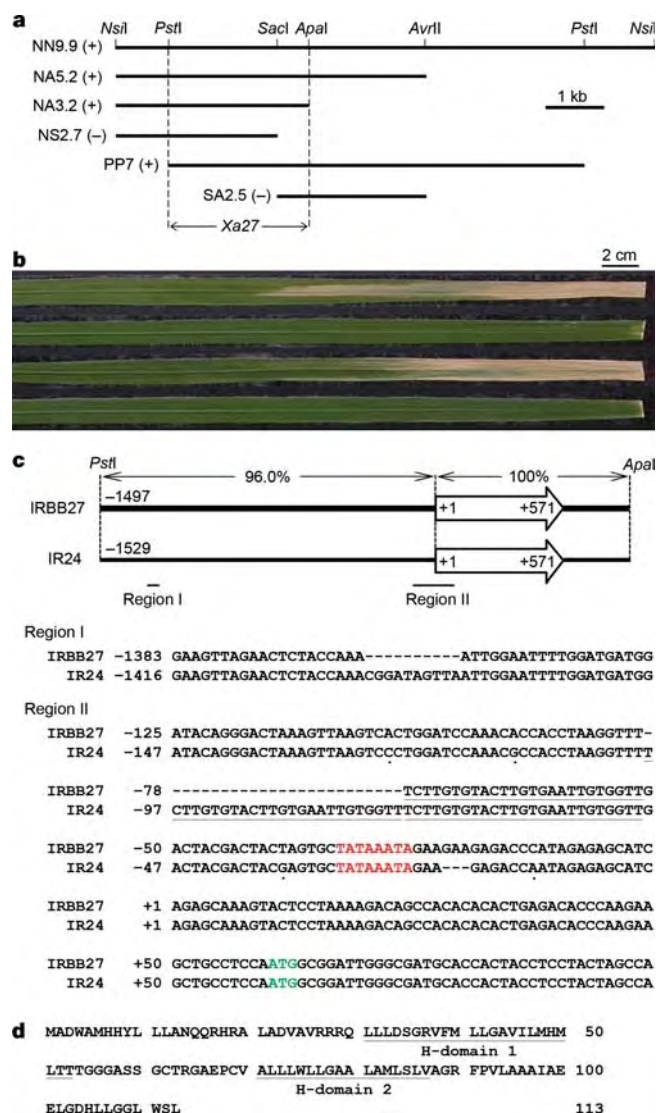


Figure 2 | Identification of *Xa27*. **a**, Clones from IRBB7 tested for *Xa27* activity. **b**, Transgenic plant TN8 conferred resistance to PXO99^A (second leaf from top) and PXO99^AME1 (*avrXa27*) (bottom leaf), but not to PXO99^AME1 (third leaf from top). Nipponbare parent was susceptible to PXO99^AME1 with *avrXa27* (top leaf). **c**, Genomic organization and sequence alignment of the polymorphic promoter regions (regions I and II) of the *Xa27* alleles. Open arrow, predicted mRNA. The 25-bp elements are underlined. TATA boxes (red) and start codons (green) are indicated. **d**, Predicted amino-acid sequence of *Xa27* and α -helix domains (underlined).

indicating that *Xa27* resided within a 2.4-kb *PstI*-*ApaI* region (Fig. 2a and Supplementary Table 1). Challenge with bacteria resulted in resistance only if strains contained *avrXa27* (Fig. 2b). A complementary DNA was isolated from an IRBB27 cDNA library, and the 5' sequence was obtained by a 5' RACE polymerase chain reaction (PCR). *Xa27* is an intronless gene and encodes a protein of 113 amino acids (Fig. 2d). *Xa27* has no discernible sequence similarity to proteins from organisms other than rice (Supplementary Fig. 3). At least four paralogues were identified in cultivar Nipponbare, including one (BAD32948) that is allelic to *Xa27*. However, the allele in Nipponbare has diverged considerably (64% identity) from the gene of IRBB27. Structural analysis of *Xa27* provided little or no clues as to the mode-of-action of the protein. A TMPred (http://www.ch.embnet.org/software/TMPRED_form.html) analysis predicted two α -helix domains in *Xa27* (Fig. 2d). However, the significance of the predictions has not been tested.

The allele from IR24 was also characterized and shares a near-identical DNA sequence with *Xa27* from IRBB27 (Fig. 2c). Two striking differences occur in the presumed promoter. One difference is the presence of a 10-base-pair (bp) sequence at the -1.4-kb position in the susceptible allele (Fig. 2c). The second is the presence of a 25-bp duplication 18 bp upstream of the putative TATA box, also in the susceptible allele (Fig. 2c). The differences raised the possibility that the two alleles might differ in their expression. Indeed, hybridization analysis revealed that only the resistance allele was expressed,

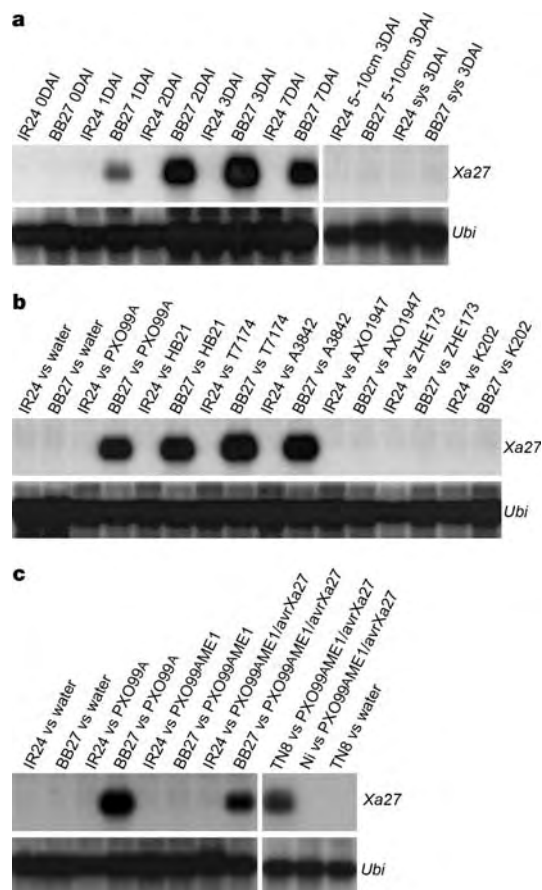


Figure 3 | *Xa27* is induced specifically during challenge by bacteria containing *avrXa27*. **a**, *Xa27* in IRBB27, but not in IR24, was induced after challenge with PXO99^A. Induction occurs in inoculated leaf tips, but not in tissue 5–10 cm below the infection sites or un-inoculated systemic leaves (sys). DAI, days after inoculation. **b**, *Xa27* was induced by incompatible strains (PXO99^A, HB21, T7174 and A3842), but not by compatible strains (AXO1947, ZHE173 and K202). **c**, *Xa27* was induced in IRBB27 or TN8 by the strains harbouring *avrXa27*, but not by the *avrXa27* mutant PXO99^AME1.

and expression of *Xa27* was only detected upon inoculation with the incompatible *X. oryzae* pv. *oryzae* strains, reaching the highest level three days after inoculation (Fig. 3a and b). *Xa27* was not induced by challenge with the mutant PXO99^ΔME1 strain, but the ability to induce *Xa27* was restored by the re-introduction of *avrXa27* (Fig. 3c). No expression of *Xa27* was detected in leaf tissue immediately below the inoculation site or from untreated leaves, indicating that induction is not systemic and is limited to the vicinity of infected tissues (Fig. 3a). No expression was detected in the susceptible plants or in *Xa27* plants challenged with compatible strains of the pathogen (Fig. 3a and b).

Expression of *Xa27* in IRBB27 rice appears to be tightly controlled as might be expected for a potentially lethal effect upon expression. However, several *Xa27* transgenic lines, including TT41, showed resistance to compatible as well as incompatible strains, indicating that these lines might have increased expression levels of *Xa27*, possibly due to the location of the T-DNA insert (Supplementary Table 2). TT41 plants, although normal in outward appearance, showed thickened vascular elements in the absence of bacterial challenge, which are typical of plants such as IRBB27 after challenge with incompatible bacteria (Fig. 4a). Unlike IRBB27, *Xa27* expression was detected in unchallenged TT41 plants (Fig. 4b). The phenotype of TT41 indicated that the expression of *Xa27* is a critical feature for the resistance reaction. To further examine the consequence of ectopic expression, the *Xa27* coding sequence was placed under the control of the rice *PR1* promoter, which is not activated specifically by *AvrXa27*. However, *PR1* is induced in a variety of conditions including infection by compatible and incompatible bacteria (unpublished data). Six independent *P_{PR1}Xa27* transgenic lines were examined for *Xa27* expression and resistance with similar results. The plants expressed *Xa27* in a pattern similar to the native *PR1* (Fig. 4c) and were resistant to both incompatible and compatible strains of the pathogen (Fig. 4d, Supplementary Table 2).

R genes are often expressed constitutively in uninfected plants. In a few cases, *R* gene expression is elevated upon infection^{8,15,16}. *Xa1*, for example, is a race-specific resistance gene in rice to *X. oryzae* pv. *oryzae* and induced by bacterial inoculation. However, expression of *Xa1* is increased by wounding and challenge with compatible and

incompatible bacteria⁸. In none of the above cases is specificity proposed to be the result of differential expression of the *R* gene. With a few exceptions, *Xa27* is not expressed at detectable levels in rice leaves unless it has been challenged with the incompatible bacteria and, more specifically, in the presence of *AvrXa27*. The localization of *Xa27* expression is consistent with the hypothesis that induction only occurs within those cells that receive *AvrXa27* via the type-III secretion system. Ectopic expression of *Xa27* circumvented the requirement for *AvrXa27* and conferred resistance on compatible strains, indicating that resistance is a post-transcriptional consequence of *Xa27* expression and that specificity is attributable to the *Xa27* promoter. The presence of the 10-bp sequence and the 25-bp duplication in the null allele implies that these regions may be necessary for expression that is influenced by *AvrXa27*.

Xa27 and *avrXa27*, in addition to being the first bacterial avirulence and *R*-gene combination from rice, provide the first insight into the molecular basis of pathogen recognition directed by effectors requiring nuclear localization for function^{17–19}. *AvrXa27* and the related type III effectors have features similar to transcription activator and may function as transcription factors—both as elicitors and virulence factors²⁰. The induction of *Xa27* in the presence of *AvrXa27* is consistent with the above model. However, whether *AvrXa27* acts directly at the *Xa27* promoter or indirectly through the activation of an endogenous transcription factor is unknown.

The prevailing models for specificity propose that *R*-gene products recognize elicitors or a catalytic product of an effector enzyme, and initiate resistance through a signalling cascade²¹. In models involving the *AvrRpt2* protease, for example, the plant appears to have evolved an alternate or mimic substrate to the virulence targets, which are at present unknown^{22,23}. Effectors related to *AvrXa27* are critical virulence factors, whose activities are dependent on nuclear localization and transcription activation^{11,18,24}. Indeed, we have recently identified changes in rice gene expression during compatible (disease) interactions that are due to specific *AvrXa27*-related effectors (unpublished data), and host expression changes specific to *AvrBs3* have also been observed²⁵. Specificity in the case of *Xa27* is the result of host gene induction. Here, again, the plant appears to deploy a mimic, in this case a mimic promoter, that triggers a defence response and

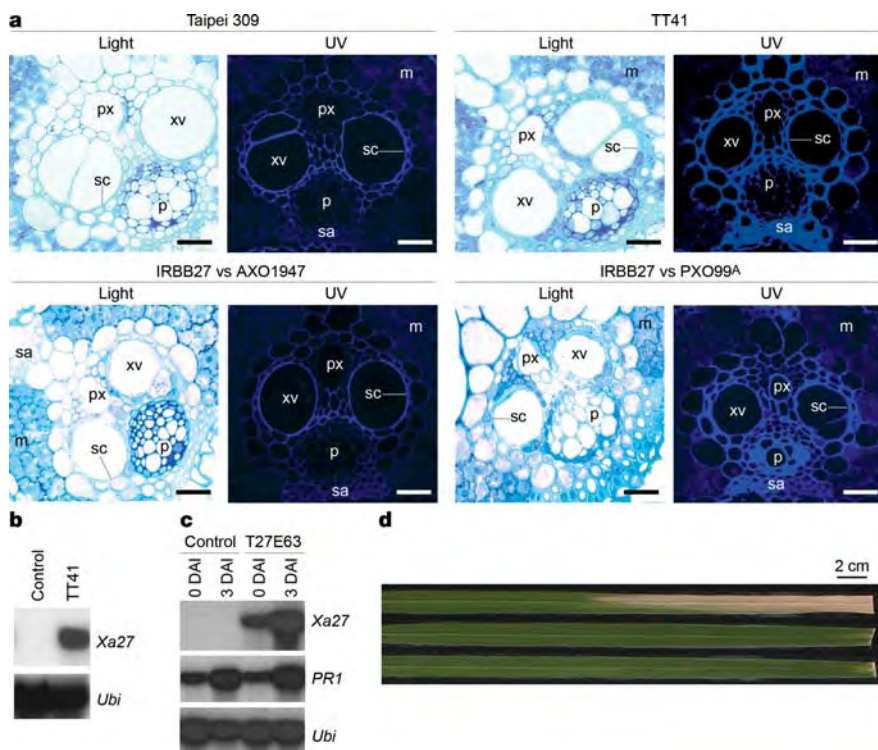


Figure 4 | Ectopic expression of *Xa27* confers resistance to compatible strains of *X. oryzae* pv. *oryzae*.

a, Secondary cell-wall thickening in vascular bundle elements of transgenic line TT41 and from IRBB27 at 3 DAI with incompatible strain PXO99^Δ. Normal elements are shown for parental (Taipei 309) plants and IRBB27 plants at 3 DAI with compatible pathogen AXO1947. Secondary cell walls appear light blue under light microscopy and fluorescent blue under confocal microscopy. Abbreviations: m, mesophyll cells; p, phloem; px, protoxylem lacuna; sa, sclerenchyma cells; sc, secondary cell wall; xv, xylem vessel; UV, ultraviolet. Scale bars, 20 μ m. **b**, Elevated expression of *Xa27* in TT41 in the absence of bacterial challenge. **c**, Expression of *P_{PR1}Xa27* in transgenic line T27E63 was elevated after inoculation with compatible strain AXO1947 (lanes 3 and 4). **d**, *P_{PR1}Xa27* line T27E63 conferred resistance to both incompatible strain PXO99^Δ (middle leaf) and compatible strain AXO1947 (bottom leaf). Control line was susceptible to compatible strain AXO1947 (top leaf). For control experiments in **b**, **c** and **d**, transgenic Taipei 309 (**b**) or Nipponbare (**c** and **d**) lines with the pC1300 vector were used.

confounds the desired effects of the pathogen. Molecular mimicry is a term sometimes used to describe pathogens that mimic host molecules to confound host recognition and defence²⁶. The evidence from plant recognition of pathogenic bacteria suggests a general model of molecular mimicry in the adaptation of plants in response to pathogens and their associated virulence strategies.

METHODS

Isolation of *avrXa27*. PXO99^AME1 was identified from a collection of mutants of PXO99^A (ref. 24). A subgenomic library of *avrXa7*-related sequences was introduced into PXO99^AME1, and strains were tested for *Xa27*-dependent reactions on six-week-old IRBB27 (*Xa27/Xa27*) plants. *avrXa27* was subcloned from 99-*avrXa27*-20. The NLS mutation was generated by replacing the *PmlI*-*HindIII* fragment of pZWavrXa27 with the corresponding fragment from pZWavrXa7M123 (ref. 11). *avrXa27TGA* was generated by exchanging the *SphI* fragment of pZWavrXa27 with that of pZWavrXa10TGA, which had a stop codon at the 1046 position¹⁰.

Isolation of *Xa27*. BAC libraries of IRBB27 and IR24 were made in pIndigoBAC-5 as described²⁷. Clones 43L18 and 109L05 were identified from IRBB27 using markers M964 and M499 (ref. 9), respectively, and sequenced (Supplementary Fig. 2). A clone containing allele of the *Xa27* gene was isolated from IR24 and sequenced. Subclones from 43L18 and 109L05 were inserted into the pC1300 vector and transferred to rice cultivars Nipponbare or Taipei 309 with *Agrobacterium tumefaciens* strain AGL1²⁸. Six-week-old T₀ transgenic plants were evaluated for resistance using the leaf-clipping method²⁹. Homozygous T₂ plants were tested for resistance specificity with four IRBB27 incompatible strains and three IRBB27 compatible strains. For ectopic expression, the *Xa27* coding region was fused to the promoter of rice *PR1* gene from cultivar Co39 (accession number U89895).

***Xa27* cDNA and northern blot analyses.** A cDNA library was constructed from PXO99^A-inoculated leaf tissue of IRBB27 in Uni-ZAP XR (Stratagene). RNA was isolated by using TRIZOL reagent (Invitrogen), and mRNA was purified with a messenger RNA kit (Qiagen). The 5' end of the *Xa27* cDNA was determined with SMART RACE cDNA Amplification Kit (Clontech). For northern hybridization, 5 µg mRNA from induced IRBB27 plants or 30 µg total RNA from *Xa27* ectopic lines were used and probed with ³²P-labelled DNA. The mRNA loading was assessed by hybridization to rice *Ubi1*.

Histological analysis. Leaf tips 3 mm long were fixed in 50 mM NaPO₄ buffer (pH 7.4), containing 2.5% glutaraldehyde. After dehydration with ethanol, samples were embedded using a Leica Historesin Embedding Kit. The 3-µm sections were stained with 0.05% toluidine blue O and viewed with a Leica DMRB microscope. Unstained sections were examined for autofluorescence from phenolic compounds by confocal microscopy (Zeiss LSM 510) at 405 nm with a band pass 420–480 nm emission filter.

Received 2 February; accepted 6 April 2005.

- Belkadir, Y., Subramaniam, R. & Dangl, J. L. Plant disease resistance protein signalling: NBS-LRR proteins and their partners. *Curr. Opin. Plant Biol.* **7**, 391–399 (2004).
- Flor, H. H. Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.* **9**, 275–296 (1971).
- Martin, G. B., Bogdanove, A. J. & Sessa, G. Understanding the functions of plant disease resistance proteins. *Annu. Rev. Plant Biol.* **54**, 23–61 (2003).
- Greenberg, J. T. & Vinatzer, B. A. Identifying type III effectors of plant pathogens and analyzing their interaction with plant cells. *Curr. Opin. Microbiol.* **6**, 20–28 (2003).
- Iyer, A. S. & McCouch, S. R. The rice bacterial blight resistance gene *xa5* encodes a novel form of disease resistance. *Mol. Plant Microbe Interact.* **17**, 1348–1354 (2004).
- Song, W.-Y. et al. A receptor kinase-like protein encoded by the rice disease resistance gene *Xa21*. *Science* **270**, 1804–1806 (1995).
- Sun, X. et al. *Xa26*, a gene conferring resistance to *Xanthomonas oryzae* pv. *oryzae* in rice, encodes an LRR receptor kinase-like protein. *Plant J.* **37**, 517–527 (2004).
- Yoshimura, S. et al. Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proc. Natl Acad. Sci. USA* **95**, 1663–1668 (1998).
- Gu, K. et al. High-resolution genetic mapping of *Xa27(t)*, a new bacterial blight resistance gene in rice, *Oryza sativa* L. *Theor. Appl. Genet.* **108**, 800–807 (2004).
- Zhu, W., Yang, B., Chittoor, J. M., Johnson, L. B. & White, F. F. *AvrXa10* contains an acidic transcriptional activation domain in the functionally conserved C terminus. *Mol. Plant Microbe Interact.* **11**, 824–832 (1998).
- Yang, B., Zhu, W., Johnson, L. B. & White, F. F. The virulence factor *AvrXa7* of *Xanthomonas oryzae* pv. *oryzae* is a type III secretion pathway-dependent, nuclear-localized, double-stranded DNA binding protein. *Proc. Natl Acad. Sci. USA* **97**, 9807–9812 (2000).
- Schornack, S. et al. The tomato resistance protein Bs4 is a predicted non-nuclear TIR-NB-LRR protein that mediates defense responses to severely truncated derivatives of *AvrBs4* and overexpressed *AvrBs3*. *Plant J.* **37**, 46–60 (2004).
- Bonas, U., Stall, R. E. & Staskawicz, B. J. Genetic and structural characterization of the avirulence gene *avrBs3* from *Xanthomonas campestris* pv. *vesicatoria*. *Mol. Gen. Genet.* **218**, 127–136 (1989).
- Swarup, S., De Feyter, R., Brlansky, R. H. & Gabriel, D. W. A pathogenicity locus from *Xanthomonas citri* enables strains from several pathovars of *X. campestris* to elicit canker-like lesions on citrus. *Phytopathology* **81**, 802–809 (1991).
- Levy, M., Edelbaum, O. & Sela, I. Tobacco mosaic virus regulates the expression of its own resistance gene *N*. *Plant Physiol.* **135**, 2392–2397 (2004).
- Radwan, O., Mouzeyar, S., Nicolas, P. & Bouzidi, M. F. Induction of a sunflower CC-NBS-LRR resistance gene analogue during incompatible interaction with *Plasmopora halstedii*. *J. Exp. Bot.* **56**, 567–575 (2005).
- Hopkins, C. M., White, F. F., Choi, S., Guo, A. & Leach, J. E. Identification of a family of avirulence genes from *Xanthomonas oryzae* pv. *oryzae*. *Mol. Plant Microbe Interact.* **5**, 451–459 (1992).
- Szurek, B., Marois, E., Bonas, U. & Van den Ackerveken, G. Eukaryotic features of the *Xanthomonas* type III effector *AvrBs3*: Protein domains involved in transcriptional activation and the interaction with nuclear import receptors from pepper. *Plant J.* **26**, 523–534 (2001).
- Herbers, K., Conrads-Strauch, J. & Bonas, U. Race-specificity of plant resistance to bacterial spot disease determined by repetitive motifs in a bacterial avirulence protein. *Nature* **356**, 172–174 (1992).
- White, F. F., Yang, B. & Johnson, L. B. Prospects for understanding avirulence gene function. *Curr. Opin. Plant Biol.* **3**, 291–298 (2000).
- Wu, A. J., Andriotis, V. M., Durrant, M. C. & Rathjen, J. P. A patch of surface-exposed residues mediates negative regulation of immune signalling by tomato Pto kinase. *Plant Cell* **16**, 2809–2821 (2004).
- Lim, M. T. & Kunkel, B. N. The *Pseudomonas syringae* type III effector *AvrRpt2* promotes virulence independently of RIN4, a predicted virulence target in *Arabidopsis thaliana*. *Plant J.* **40**, 790–798 (2004).
- Belkadir, Y., Nimchuk, Z., Hubert, D. A., Mackey, D. & Dangl, J. L. *Arabidopsis* RIN4 negatively regulates disease resistance mediated by RPS2 and RPM1 downstream or independent of the NDR1 signal modulator and is not required for the virulence functions of bacterial type III effectors *AvrRpt2* or *AvrRpm1*. *Plant Cell* **16**, 2822–2835 (2004).
- Yang, B. & White, F. Diverse members of the *AvrBs3*/*PthA* family of type III effectors are major virulence determinants in bacterial blight disease of rice. *Mol. Plant Microbe Interact.* **17**, 1192–1200 (2004).
- Marois, E., Van den Ackerveken, G. & Bonas, U. The *Xanthomonas* type III effector protein *AvrBs3* modulates plant gene expression and induces cell hypertrophy in the susceptible host. *Mol. Plant Microbe Interact.* **15**, 637–646 (2002).
- Stebbins, E. C. & Galán, J. A. Structural mimicry in bacterial virulence. *Nature* **412**, 701–705 (2001).
- Wang, G.-L., Holsten, T. E., Song, W.-L., Wang, H.-P. & Ronald, P. C. Construction of a rice bacterial artificial chromosome library and identification of clones linked to the *Xa-21* disease resistance locus. *Plant J.* **7**, 525–533 (1995).
- Yin, Z. & Wang, G.-L. Evidence of multiple complex patterns of T-DNA integration into the rice genome. *Theor. Appl. Genet.* **100**, 461–470 (2000).
- Kauffman, H. E., Reddy, A. P. K., Hsieh, S. P. Y. & Merca, S. D. An improved technique for evaluating resistance to rice varieties of *Xanthomonas oryzae* pv. *oryzae*. *Plant Dis. Rep.* **57**, 537–541 (1973).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Y. Teo for technical assistance, G. S. Khush, H. Leung and R. Nelson for *Xa27* germplasm, Q. Zhang and J. E. Leach for *X. oryzae* pv. *oryzae* strains, K. Shimamoto for *PR1* cDNA, CAMBIA for the pC1300 vector. This work was supported by the NRI programme of the US Department of Agriculture (F.F.W.); the Kansas Agriculture Experiment Station (B.Y. and F.F.W.); the intramural research funds from Temasek Life Sciences Laboratory (Z.Y.); a competitive grant from the Agri-Food and Veterinary Authority of Singapore (Z.Y.).

Author Contributions K.G., B.Y., D.T. and L.W. are all co-first authors. G.-L.W. initiated the preliminary studies of *Xa27*. Z.Y. designed and carried out data analysis. K.G., D.T. and L.W. conceived the experiment, and together with D.W., C.S., F.Y., Z.C. and Z.Y. carried it out. For *avrXa27*, B.Y., F.F.W. and Z.Y. designed and carried out data analysis. B.Y., together with D.T. and K.G., conducted the experiment. F.F.W. and Z.Y. co-wrote the paper.

Author Information Sequences of *O. sativa* *Xa27* IR24 allele (*xa27*), *Xa27* IRBB27 allele (*Xa27*), *Xa27* full-length cDNA and *X. oryzae* pv. *oryzae* PXO99^A *avrXa27* locus are deposited in GenBank under accession numbers AY986491, AY986492, AY986493 and AY986494. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to Z.Y. (yinzc@tll.org.sg).

LETTERS

EphB receptor activity suppresses colorectal cancer progression

Eduard Batlle^{1,2,3}, Julinor Bacani⁵, Harry Begthel¹, Suzanne Jonkeer^{1,2}, Alexander Gregorieff¹, Maaike van de Born¹, Núria Malats⁶, Elena Sancho^{1,2}, Elles Boon⁴, Tony Pawson⁵, Steven Gallinger⁵, Steven Pals⁴ & Hans Clevers¹

Most sporadic colorectal cancers are initiated by activating Wnt pathway mutations¹, characterized by the stabilization of β -catenin and constitutive transcription by the β -catenin/Tcf4 complex^{2,3}. EphB guidance receptors are Tcf4 target genes that control intestinal epithelial architecture through repulsive interactions with Ephrin-B ligands^{4,5}. Here we show that, although Wnt signalling remains constitutively active, most human colorectal cancers lose expression of EphB at the adenoma–carcinoma transition. Loss of EphB expression strongly correlates with degree of malignancy. Furthermore, reduction of EphB activity accelerates tumorigenesis in the colon and rectum of *Apc*^{Min/+} mice, and results in the formation of aggressive adenocarcinomas. Our data demonstrate that loss of EphB expression represents a critical step in colorectal cancer progression.

The genetic programme driven by β -catenin and Tcf-4 in colorectal cancer (CRC) closely resembles that of epithelial proliferative progenitors in the intestinal crypts³. As a case in point, EphB2 outlines epithelial cells in normal crypts, but is also expressed in the earliest neoplastic lesions caused by a mutationally activated Wnt pathway (Fig. 1a)^{4–6}. Mutations in APC or β -catenin are present in the great majority of CRCs^{7–9}. Consequently, constitutive β -catenin/Tcf-4 activity is present in virtually all CRC cell lines^{2,3}. Unexpectedly, we found that *EPHB2* messenger RNA expression was greatly reduced in most CRC cell lines compared to our reference CRC cell line Ls174T (Fig. 1b). This finding was confirmed in a small panel of human CRC samples, including lymph node and liver metastases (examples within Fig. 1 and Supplementary Fig. 1). EphB2 could be observed in the progenitor cells at the bottom of crypts in adjacent normal tissue (Fig. 1d). All adenocarcinomas showed prominent nuclear β -catenin accumulation (Fig. 1c, e). Yet, nine out of ten showed extensive areas of highly reduced or completely absent EphB2 immunostaining (Fig. 1d, f). In metastases, colonizing CRC cells accumulated β -catenin, yet in five out of six cases EphB2 expression was absent or strongly diminished (Supplementary Fig. 1). We concluded that secondary silencing of EphB2 expression occurs frequently in malignant CRCs independently of β -catenin/Tcf-4 activity.

To determine at what stage of the adenoma–carcinoma progression EphB2 expression was lost in CRCs, we stained 30 dysplastic aberrant crypt foci (dACF), 31 small adenomas (diameter <5 mm), 12 large adenomas (diameter $\geq$ 5 mm) and 36 carcinomas of different Dukes stages. The percentage of EphB2-positive cells was assessed and lesions were grouped. Group I lesions contained more than 80% EphB2-positive cells. Group II lesions contained 50–80% EphB2-positive cells. Group III lesions contained 20–50% positive cells,

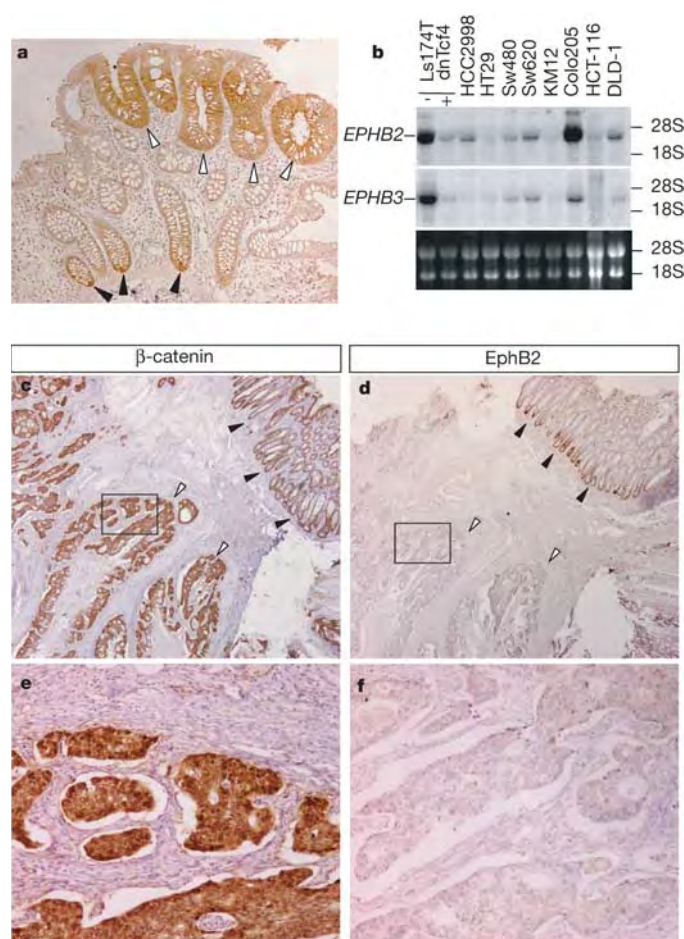


Figure 1 | *EPHB2* is a β -catenin/Tcf-4 target gene but it is downregulated in advanced colorectal tumours and cell lines. **a**, EphB2 immunostaining of an early colorectal lesion from a Familial Adenomatous Polyposis patient. **b**, Analysis of *EPHB2* and *EPHB3* mRNA levels in a panel of CRC cell lines by northern blot. First two lanes correspond to Ls174T cells before (–) and after (+) expression of a dominant negative form of Tcf-4 (ref. 5). Bottom panel shows RNA loading. 28S and 18S indicate the migration of ribosomal subunits. **c–f**, Immunostaining using anti-EphB2 (**d, f**) or anti- β -catenin antibodies (**c, e**) of an invasive colorectal adenocarcinoma. Panels **e** and **f** show high magnification pictures of the fields labelled with boxes in panels **c** and **d**. Black arrowheads point to EphB2-positive progenitor cells at the crypt base. White arrowheads point to tumour cells.

¹Hubrecht Laboratory, Center for Biomedical Genetics, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands. ²Biomedical Research Institute, Barcelona Science Park, Josep Samitier 1-5, 08028 Barcelona, Spain. ³Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig de Lluís Companys 23, 08010 Barcelona, Spain. ⁴Department of Pathology, Academic Medical Center, 1105 AZ Amsterdam, The Netherlands. ⁵Samuel Lunefeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario, M5G 1X5, Canada. ⁶Institut Municipal d'Investigació Mèdica, Dr Aiguader 80, 08003 Barcelona, Spain.

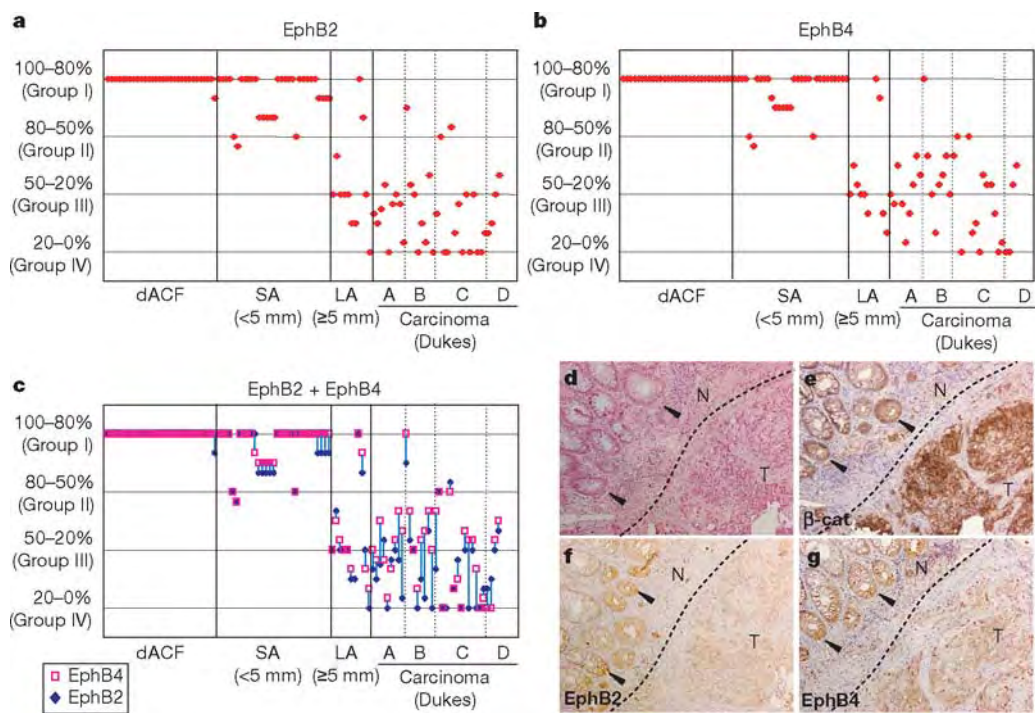


Figure 2 | EphB2 and EphB4 are downregulated during CRC progression. **a, b**, Classification of 108 colorectal lesions from patients according to the percentage of EphB2-positive (**a**) or EphB4-positive (**b**) cells and tumour stage. Each lesion is depicted as a red diamond. Dysplastic aberrant crypt foci (dACF), small adenomas (SA), large adenomas (LA) and carcinomas differed significantly between each other regarding EphB2 or EphB4 levels ($P = 0.001$; Supplementary Figs 3 and 4). **c**, A strong overall correlation

($r = 0.964$, $P < 0.001$; Supplementary Fig. 5) was observed between the percentage of EphB2-positive (blue diamonds) and EphB4-positive (magenta squares) cells. A line links both scores in each lesion. **d–g**, Example of coordinated silencing of EphB2 (**f**) and EphB4 (**g**) in a carcinoma showing nuclear β -catenin accumulation (**e**). Dashed line depicts the boundary between tumour (T) and normal tissue (N). Arrowheads indicate staining in colonic crypts of healthy tissue.

whereas lesions showing less than 20% positive cells were classified in group IV. Staining in adjacent normal crypts was always used as an internal reference. Examples for each EphB2 group are shown in Supplementary Fig. 2. The mean score and stage for each lesion are depicted as a scatter plot in Fig. 2a. All (30 of 30) dACFs and 61% (19 of 31) of small adenomas retained EphB2 expression. Virtually all adenocarcinomas showed extensive loss of EphB2 expression (>50% of negative cells), while 25% of these malignant tumours were entirely EphB2-negative.

A strong association was found between histological tumour grade and EphB2 silencing independent of Dukes staging (Fig. 3a). Of 19 high grade adenocarcinomas, 14 completely lacked EphB2 expression and were classified in Group IV while the rest showed less than 50% EphB2-positive cells. The difference in EphB2 reactivity between high grade and low-intermediate grade adenocarcinomas had strong statistical significance ($P = 0.017$; Mann-Whitney). In Group II/III, tumour masses were commonly composed of

intermingled EphB2-positive and -negative cells (Supplementary Fig. 2). However, some tumours contained areas that stained largely positive, adjacent to areas with downregulated EphB2 levels (Fig. 3a; cases linked by a dashed line). In these latter cases, the negative areas (Fig. 3f) were invariably high grade (Fig. 3e). *In situ* hybridization combined with immunohistochemistry on sequential sections demonstrated that in 12 out of 15 tumours analysed, *EPHB2* mRNA strictly followed protein expression (examples within Fig. 3b–g). The remaining three adenocarcinomas showed evident protein downregulation yet *EPHB2* mRNA remained present (Supplementary Fig. 6).

EphB3, the closest homologue of EphB2, was also silenced in CRC cell lines (Fig. 1b). Unfortunately, our antibodies did not allow consistent analysis of EphB3 expression in human tissue. However, by *in situ* hybridization on the above panel of 15 carcinomas, *EPHB3* mRNA was found to be downregulated with virtually identical patterns to that of *EPHB2* (Supplementary Fig. 7). Extending these

Table 1 | CRC progression in *Apc*^{Min/+} mice upon loss of *Ephb3*

	<i>Ephb3</i> ^{+/+} ; <i>Apc</i> ^{Min/+}	<i>Ephb3</i> ^{+/-} ; <i>Apc</i> ^{Min/+}	<i>Ephb3</i> ^{-/-} ; <i>Apc</i> ^{Min/+}
Colorectal tumours* (<i>n</i> = number of mice)	11 ± 6.15 (<i>n</i> = 25)	15.67 ± 11.611 (<i>n</i> = 55)	19.7 ± 12.9 (<i>n</i> = 20)
Diameter (<i>n</i> = number of tumours sized)	(<i>n</i> = 218)	(<i>n</i> = 306)	(<i>n</i> = 189)
<1 mm	48% (105)	42% (128)	44% (83)
1–5 mm	48% (104)	50% (153)	36% (69)
>5 mm†	4% (9)	8% (25)	20% (36)
Percentage of mice bearing invasive carcinomas‡	0% (<i>n</i> = 19)	16% (<i>n</i> = 37)	47% (<i>n</i> = 17)

* The mean number of colon polyps in the *Ephb3*^{-/-};*Apc*^{Min/+} group is significantly different to that of the *Ephb3*^{+/+};*Apc*^{Min/+} group ($P = 0.015$, *t*-test; $P = 0.042$, Mann-Whitney).
† The mean number of large (>5 mm) colorectal tumours in *Ephb3*^{-/-};*Apc*^{Min/+} mice is significantly different to that of *Ephb3*^{+/+};*Apc*^{Min/+} littermates ($P = 0.004$, *t*-test; $P = 0.001$, Mann-Whitney), and to that of heterozygous *Ephb3*^{+/-};*Apc*^{Min/+} ($P = 0.015$, *t*-test; $P = 0.006$, Mann-Whitney). Yet, no significant differences were found regarding small (<1 mm) and medium size (1–5 mm) tumours between any of the groups. The three genotypes significantly differ regarding the number of large colorectal polyps ($P = 0.002$, Kruskal-Wallis test).
‡ The complete loss of *Ephb3* is significantly correlated with the presence of invasion relative to *Ephb3*^{+/+};*Apc*^{Min/+} ($P = 0.001$, Fisher's exact test) or to *Ephb3*^{+/-};*Apc*^{Min/+} ($P = 0.023$, Fisher's exact test). The loss of one *Ephb3* allele was also sufficient to produce the invasive phenotype in six mice, but was not significantly correlated with invasion relative to wild-type littermates ($P = 0.086$, Fisher's exact test).

observations beyond our original array data⁵ to all *EPHA* and *EPHB* genes, we found that *EPHB4* was also a Tcf target gene as its expression in CRC cell lines was inhibited upon blocking the Wnt cascade (Supplementary Fig. 8). Indeed, EphB4 was expressed in human colonic crypts and in early CRC lesions (Supplementary Fig. 9). Again its expression was lost in advanced colorectal tumours (Fig. 2b, Supplementary Fig. 9). Individual CRC samples showed similar percentages of EphB2- and EphB4-negative cells (Fig. 2c, Supplementary Fig. 5) and overlapping expression domains within each tumour (examples within Fig. 2d–g) implying that expression of these genes was silenced coordinately during CRC progression.

Although CRCs appear under strong selective pressure to down-regulate EphB2, EphB3 and EphB4, our observations did not prove a causal role for EphB silencing in tumour progression. To test this, we crossed the *Apc*^{Min} allele into a transgenic mouse line that expresses in the intestinal epithelium a dominant-negative form of EphB2 lacking the cytoplasmic tail (Δ^{cy} EphB2)⁴. The membrane anchored Δ^{cy} EphB2 molecule competes with EphB2, -B3 and -B4 receptors for binding of common Ephrin-B ligands, but is unable to transduce signals. Indeed, Δ^{cy} Ephb2 transgenic mice phenocopy the individual intestinal phenotypes caused by disruption of *Ephb2* (defective cell positioning along the crypt-axis) and of *Ephb3* (loss of directional

sorting of Paneth cells)⁴, implying that the transgene faithfully mimics *Ephb* loss-of-function mutations. We never observed intestinal tumours in Δ^{cy} Ephb2 transgenics ($n > 20$; ages > 6 months).

Apc^{Min/+} mice bear a truncated *Apc* allele¹⁰ and develop dozens of adenomas in the small intestine¹¹ with accumulated nuclear β -catenin and high expression of EphB2 (ref. 4), EphB3 and EphB4 (Supplementary Fig. 8). Of note, unlike in humans carrying *Apc* mutations, macroscopic colorectal tumours in *Apc*^{Min/+} mice are infrequent. Yet, multiple small dysplastic crypts, the earliest precursors of CRC, have been described in the colon of *Apc*^{Min/+} mice¹². We confirmed these observations. *Apc*^{Min/+} mice showed only 0.5 ± 0.5 macroscopic (diameter > 1 mm) colorectal tumours, but developed multiple colonic microlesions (9 ± 4), the majority ($> 85\%$) arising in the distal third of the large intestine ($n = 5$ at 21 weeks of age).

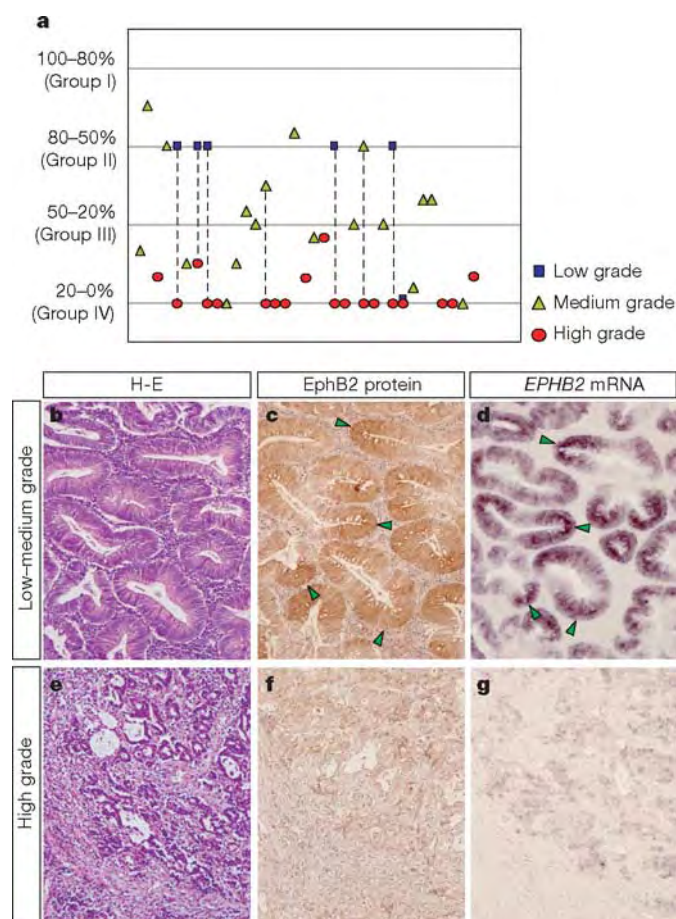


Figure 3 | EphB2 downregulation correlates with higher histological tumour grade. **a**, Correlation between the percentage of EphB2-positive cells and tumour grading in carcinomas. Areas of different grade within the same tumour are represented linked by a dashed line. **b–g**, Example of a single carcinoma showing a low-medium grade component (**b–d**) adjacent to a high-grade area (**e–g**). Consecutive sections were stained using anti-EphB2 antibodies (**c, f**) or by *in situ* hybridization with an *EPHB2* antisense cRNA probe (**d, g**). Arrowheads point to tumour glands that stained strongly positive for EphB2 protein and mRNA in the low-grade area. Panels **b** and **e** are haematoxylin-eosin (H-E) stainings.

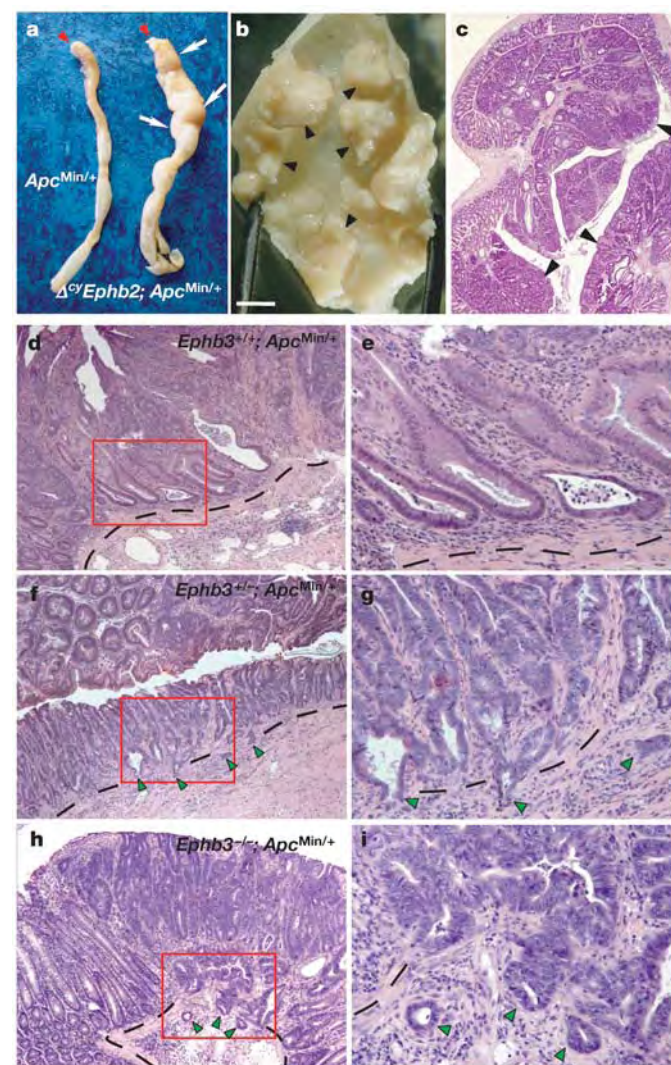


Figure 4 | Accelerated colorectal tumorigenesis in *Apc*^{Min/+} mice expressing Δ^{cy} EphB2 transgene or bearing *Ephb3*-null alleles. **a**, Macroscopic comparison of the colon of *Apc*^{Min/+} and Δ^{cy} Ephb2;*Apc*^{Min/+} mice at 21 weeks of age. Arrows point to distended areas in Δ^{cy} Ephb2;*Apc*^{Min/+} mice. Red arrowheads indicate the distal end. **b**, A dissected colorectum of a Δ^{cy} Ephb2;*Apc*^{Min/+} mouse showing multiple tumours (arrowheads). Scale bar, 5 mm. **c**, Colon histology in compound Δ^{cy} Ephb2;*Apc*^{Min/+} mice. **d–i**, Haematoxylin-eosin staining of representative colorectal tumours of *Ephb3*^{+/+};*Apc*^{Min/+} (**d, e**), *Ephb3*^{+/-};*Apc*^{Min/+} (**f, g**) and *Ephb3*^{-/-};*Apc*^{Min/+} (**h, i**) mice. Dashed line depicts the muscularis mucosae. Green arrows point to invasive glands. Panels **e, g** and **i** show higher magnification pictures of the fields labelled with boxes in panels **d, f** and **h**, respectively.

Colonic microlesions in $Apc^{Min/+}$ mice also accumulated high levels of β -catenin and expressed EphB2, EphB4 (Supplementary Fig. 10) and EphB3 (data not shown). Thus, although tumorigenesis is initiated frequently in the colorectum of $Apc^{Min/+}$ animals, tumours do not progress beyond the earliest stage.

Compound $\Delta^{cy}Ephb2;Apc^{Min/+}$ animals became moribund at the same age as control $Apc^{Min/+}$ mice (~23 weeks). Post-mortem examination revealed decreased tumour counts in the small intestine of compound $\Delta^{cy}Ephb2;Apc^{Min/+}$ (13 ± 4 macroscopic tumours per mouse compared to 39 ± 12 tumours in control $Apc^{Min/+}$ mice). Strikingly, all $\Delta^{cy}Ephb2;Apc^{Min/+}$ mice analysed showed distension of the colorectum, unprecedented in $Apc^{Min/+}$ mice (Fig. 4a). Dissection revealed 11 ± 3 macroscopic tumours ($n = 7$ at 21 weeks of age) clustered in the distal colon and rectum (Fig. 4b) compared with a single macroscopic tumour per mouse found in the distal colon of $Apc^{Min/+}$ non-transgenic littermates ($n = 2$). $\Delta^{cy}Ephb2;Apc^{Min/+}$ colorectal tumours grew as large polypoid structures (mean diameter 4.5 ± 1.5 mm). They were highly dysplastic with profound architectural and cytological distortion to a level that is highly uncommon in human or murine polyps but characteristic of carcinomas (Fig. 4c). In addition, we frequently observed pronounced cribriform tumour growth with areas of necrosis and prominent desmoplastic stromal reaction suggesting invasion of the lamina propria (Supplementary Fig. 11). Based on this, all $\Delta^{cy}Ephb2;Apc^{Min/+}$ tumours were unequivocally classified as intra-mucosal adenocarcinomas (>30 tumours from 7 different mice), indicating a fully penetrant phenotype.

We also obtained $Apc^{Min/+}$ mice carrying *Ephb3* null alleles (Table 1). Control $Ephb3^{+/+};Apc^{Min/+}$ littermates developed significant numbers of colorectal polyps, presumably owing to the mixed genetic background or differences in mice stocks (see Methods section). Nevertheless, loss of EphB3 accelerated tumorigenesis. $Ephb3^{+/-};Apc^{Min/+}$ and $Ephb3^{-/-};Apc^{Min/+}$ mice consistently showed higher numbers and larger colorectal polyps than control littermates. 20% of the neoplasms in $Ephb3^{-/-};Apc^{Min/+}$ were very large adenocarcinomas (diameter >5 mm) that rarely arose in $Ephb3^{+/+};Apc^{Min/+}$ littermates. Furthermore, around half of the $Ephb3^{-/-};Apc^{Min/+}$ animals developed carcinomas that invaded the muscle layer, a feature of malignancy that was not present in $Ephb3^{+/+};Apc^{Min/+}$ tumours (Fig. 4d–i). Even a single *Ephb3* null allele enhanced invasive behaviour of $Apc^{Min/+}$ tumours (Fig. 4f, g). Overall, these observations confirmed that reduction of EphB activity exacerbated colorectal tumorigenesis in $Apc^{Min/+}$ mice and provided further evidence for a causal role of EphB silencing in CRC progression.

In the large intestine, dysplastic crypts and small polyps remain confined to small areas at the surface epithelium (Supplementary Fig. 10) where they are surrounded by normal cells expressing high levels of ephrin-B ligands (data not shown). Our observations indicate that EphB activity in CRC cells prevents further expansion and malignant progression of these benign lesions. We propose that human colorectal tumours overcome the restriction imposed by EphB receptors by silencing their expression. It has been recently reported that EphB2 becomes mutationally inactivated in a significant fraction of prostate tumours, thus reinforcing the idea that EphB activity acts as a tumour suppressor¹³. The mechanism behind the coordinated EphB downregulation in CRC is currently unknown. Our data indicate that silencing in CRC occurs at transcriptional or mRNA level, yet alterations only affecting protein levels are present in a subset of colorectal carcinomas.

We recently distinguished two different physiological functions for the β -catenin/Tcf-4 target gene programme in the intestinal epithelium: first, the maintenance of the undifferentiated, proliferative crypt phenotype, and second, the control of cell positioning along the crypt–villus axis through expression of EphB receptors^{4,5,14}. The first function is maintained throughout all stages of carcinogenic progression⁵, presumably because it dictates an essential feature of

the transformed phenotype. During tumour progression, selection can occur against the expression of certain β -catenin/Tcf target genes, such as the *EPHB* genes. Our data illustrate the complexity of the mechanisms operating during tumour progression where initial mutational activation of a pathway confers certain selective advantages for tumour growth, but simultaneously imposes restrictions on immediate tumour progression. EphB2 has been proposed as a target for antibody-based cancer therapy on the basis of its relative overall upregulation in CRC compared to normal intestinal tissue¹⁵. This apparent discrepancy with our data might be explained by the fact that EphB2 expression in healthy intestine is restricted to less than 10% of all cells (that is, to a few progenitors at the crypt base). Thus, even carcinomas composed of only 25% EphB2-positive cells will show at least twofold higher overall EphB2 expression compared with normal tissue. Our current data warrant caution for therapeutic strategies, and underscore the necessity for careful, extensive validation of potential drug targets.

METHODS

Cell lines and northern blots. Cell lines were obtained from the ATCC or the NCI. Cells were plated at low density ($25,000 \text{ cells cm}^{-2}$) and cultured in RPMI 10% FCS for 24 h before harvesting. Ls174T cells expressing dominant negative Tcf4 upon addition of doxycycline were previously described⁶. Northern blot analysis of EphB2 was described elsewhere⁴.

Antibodies and immunostaining. The anti-EphB2-antibody was a goat affinity purified polyclonal directed against the extracellular domain (R & D Systems). We have previously characterized the specificity of this antibody⁴. Anti-EphB4 was a rabbit polyclonal generated against the 50 carboxy-terminal amino acids of the cytoplasmic tail (a gift from A. Ziemiecki, University of Bern)¹⁶. Anti- β -catenin antibody was a mouse monoclonal generated against the C-terminal domain (Transduction Labs). Immunostaining methods are described in detail as Supplementary Information or elsewhere^{4,5}.

Sample collection and tumour scoring. Paraffin-embedded tissues were obtained from the files of the Department of Pathology, Academic Medical Center, University of Amsterdam, The Netherlands. Adenomas were classified as small (diameter <5 mm) or large (diameter >5 mm). Colorectal carcinomas were staged according to the original Dukes classification; Dukes A, disease limited to the bowel wall; Dukes B, extensions through the deep muscle without metastases; and Dukes C/D, tumours with regional and/or distant metastases, respectively. Tumours were graded (high, medium and low) using standard histopathological criteria.

Each sample was classified by comparing the staining in the tumour with that of the bottom of the crypts in adjacent normal tissue. Tumour samples without normal tissue were not included in the study. Neither were samples of liver or lymph node metastases included, as they did not contain intestinal epithelial tissue as reference. Three independent observers (E.B., E.B. and S.P.) subjectively assessed the percentage of EphB2- or EphB4-positive cells within the tumour and classified them in one of the four groups detailed in the text. The final score is the mean of the classifications given by the three observers. In all lesions, standard deviation was always less than 1. Greater deviations corresponded to groups II and III while virtually all lesions completely positive or completely negative were given the same score by all observers. Tumour areas still showing perceptible EphB2 or EphB4 were evaluated as negative when intensity was severely reduced compared to the staining at the bottom of the crypts in adjacent normal tissue (that is, at least fourfold reduction).

In situ hybridization. *In situ* hybridization on formalin fixed-paraffin embedded pathological material was performed as previously described¹⁷. The protocol is described in detail in Supplementary Methods. DIG-labelled antisense complementary RNA probes corresponded to 3' end mRNA of human *EPHB2* (nucleotides 3682–4153; Human Gene Index THC2025076) or of human *EPHB3* (nucleotides 3530–4251; Human Gene Index THC2016077).

Mice. Transgenic mice expressing dominant negative EphB2 in the intestine under the control of the Villin-promoter were obtained by microinjection of single cell C57BL/6J $\times$ CBA/J hybrid embryos as described previously⁴. Transgenic animals were crossed to the C57BL/6 strain for at least six generations before their mating with $Apc^{Min/+}$ mice. The generation of *Ephb3*-null mice was previously described¹⁸. *Ephb3* mice were maintained in a mixed 129/sv:C57BL/6J (1:1) outbred background. All stocks of $Apc^{Min/+}$ mice (Jackson Laboratory) were maintained in a homogenous inbred C57BL/6 background ($N > 10$).

All experiments using $\Delta^{cy}Ephb2$ mice were performed in the animal facility of Utrecht University, while *Ephb3* experiments were conducted in the facilities of Samuel Lunenfeld Research Institute. Different C57BL/6J; $Apc^{Min/+}$ mouse

stocks were used in both cases. Despite identical genetic background, we noticed that *Apc*^{Min/+} mice housed in Samuel Lunenfeld Research Institute facility developed more colorectal polyps (on average, 5 macroscopic tumours per mouse at 5 months of age) than *Apc*^{Min/+} housed in University of Utrecht facility (on average, 0.5 macroscopic tumours per mouse at 5 months of age). Both numbers are within the range of colorectal tumours reported previously for C57BL/6;*Apc*^{Min/+} mice housed in these facilities^{19,20} as well as in other laboratories^{11,12} and probably reflect differences in the diet or husbandry conditions. All animal experiments were approved by the Animal Care and Use Committees of Utrecht University and of the Samuel Lunenfeld Research Institute.

Tumour analysis in *Ephb3;Apc*^{Min/+} mice. At post-mortem examination, murine intestine was removed, opened along the longitudinal axis and fixed flat in 10% buffered formalin for 24 h at room temperature. Fixed intestines were stained with 0.5% methylene blue in distilled water for easier identification of small tumours, such as aberrant crypt foci (ACF). All colorectal lesions were identified under the dissection lens, counted and measured. Representative colon polyps were harvested, embedded in paraffin and sectioned at 5 µm. Histopathology was evaluated using standard techniques. True invasion through the muscle layer was determined by evaluating a minimum of three to six serial sections.

Received 9 February; accepted 11 April 2005.

1. Bienz, M. & Clevers, H. Linking colorectal cancer to Wnt signaling. *Cell* **103**, 311–320 (2000).
2. Korinek, V. *et al.* Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC^{-/-} colon carcinoma. *Science* **275**, 1784–1787 (1997).
3. Morin, P. J. *et al.* Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* **275**, 1787–1790 (1997).
4. Batlle, E. *et al.* Beta-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell* **111**, 251–263 (2002).
5. van de Wetering, M. *et al.* The β-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* **111**, 241–250 (2002).
6. Kuhnert, F. *et al.* Essential requirement for Wnt signaling in proliferation of adult small intestine and colon revealed by adenoviral expression of Dickkopf-1. *Proc. Natl Acad. Sci. USA* **101**, 266–271 (2004).
7. Powell, S. M. *et al.* APC mutations occur early during colorectal tumorigenesis. *Nature* **359**, 235–237 (1992).
8. Sparks, A. B., Morin, P. J., Vogelstein, B. & Kinzler, K. W. Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer. *Cancer Res.* **58**, 1130–1134 (1998).
9. Shitoh, K. *et al.* Frequent activation of the beta-catenin-Tcf signaling pathway in nonfamilial colorectal carcinomas with microsatellite instability. *Genes Chromosom. Cancer* **30**, 32–37 (2001).
10. Su, L. K. *et al.* Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene. *Science* **256**, 668–670 (1992).
11. Moser, A. R., Pitot, H. C. & Dove, W. F. A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. *Science* **247**, 322–324 (1990).
12. Yamada, Y. *et al.* Microadenomatous lesions involving loss of Apc heterozygosity in the colon of adult Apc(Min/+) mice. *Cancer Res.* **62**, 6367–6370 (2002).
13. Huusko, P. *et al.* Nonsense-mediated decay microarray analysis identifies mutations of *EPHB2* in human prostate cancer. *Nature Genet.* **36**, 979–983 (2004).
14. Sancho, E., Batlle, E. & Clevers, H. Live and let die in the intestinal epithelium. *Curr. Opin. Cell Biol.* **15**, 763–770 (2003).
15. Mao, W. *et al.* EphB2 as a therapeutic antibody drug target for the treatment of colorectal cancer. *Cancer Res.* **64**, 781–788 (2004).
16. Munarini, N. *et al.* Altered mammary epithelial development, pattern formation and involution in transgenic mice expressing the EphB4 receptor tyrosine kinase. *J. Cell Sci.* **115**, 25–37 (2002).
17. Gregorieff, A., Grosschedl, R. & Clevers, H. Hindgut defects and transformation of the gastro-intestinal tract in Tcf4(–/–)/Tcf1(–/–) embryos. *EMBO J.* **23**, 1825–1833 (2004).
18. Orioli, D., Henkemeyer, M., Lemke, G., Klein, R. & Pawson, T. Sek4 and Nuk receptors cooperate in guidance of commissural axons and in palate formation. *EMBO J.* **15**, 6035–6049 (1996).
19. Roose, J. *et al.* Synergy between tumour suppressor APC and the beta-catenin-Tcf4 target Tcf1. *Science* **285**, 1923–1926 (1999).
20. Lal, G. *et al.* Suppression of intestinal polyps in Msh2-deficient and non-Msh2-deficient multiple intestinal neoplasia mice by a specific cyclooxygenase-2 inhibitor and by a dual cyclooxygenase-1/2 inhibitor. *Cancer Res.* **61**, 6131–6136 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Ziemiecki for EphB4 antiserum, and A. García de Herreros and C. Francí for help with mouse experiments.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.C. (clevers@niob.knaw.nl).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Fixing a broken record

During a recent holiday in northern California, I met a young scientist who was unsure about her graduate-school options. As an undergraduate, she worked in a really good developmental-biology laboratory, was interested in neuroscience, but was having difficulty getting a graduate position — despite some early publications and affiliation with a star of the field.

At this point, I slipped into a 'broken record' riff I've used at career fairs, round-tables and casual conversations this year. I told her to think of all possible career opportunities as a matrix. The vertical columns represent sectors and include academia, government and industry. The horizontal rows represent the broad disciplines of life sciences and physical sciences, with chemistry acting as an interdisciplinary interface between the two. Each 'cell' is divided into 'on the bench' and 'off the bench' jobs.

In this young scientist's case, for example, she might consider emphasizing biochemistry as a way to work her way into neuroscience. The supply-and-demand situation is better in chemistry, and straddling the disciplines allows more options in all sectors. Or she

might treat her technician post as a way to gain more skills, then switch sectors to pick up fresh techniques, before pursuing more formal training.

After the conversation, I reminded myself again that if I can talk this talk, then *Naturejobs* should walk this walk, by helping scientists representing all disciplines. In the past five years or so, we've included content that covers much of the grid — and a good part of the globe. Now, our relaunch of www.naturejobs.com will help readers navigate content and adverts by discipline, skill level and sector, as well as by geography through an interactive map. We are pleased to offer even more services, more direct contact with readers, and more opportunities for readers to participate in *Naturejobs'* own online matrix. And personally, I'm pleased to have found a way to fix my broken record.



Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London
 The Macmillan Building, 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director:
 Nevin Bayoumi (4978)
European Sales Manager:
 Andy Douglas (4975)

UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
Natureevents: Silke Opstrup (4994)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
Germany/Austria/The Netherlands:
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
Germany/Austria/
The Netherlands:
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 345 Park Avenue South,
 10th Floor, New York,

NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Picture this

Buoyed by a range of new technologies, science illustration is expanding its remit to offer careers beyond publishing in areas such as advertising and law. **Virginia Gewin** reports.

When Gayle Gross de Núñez took a peek into a microscope during a mandatory lab class, she changed the course of her career. She turned her back on her liberal-arts degree to learn more about the science behind the miniature universe that first beguiled her. With her scientific curiosity satisfied, she later returned to her arts background to combine both of her passions to pursue a career illustrating the microscopic world (see 'Painting the brain', below).

Scientific illustration — once relegated to a few niche jobs at a handful of publishing houses and museums — is now evolving and growing. New tools in computer-based design and illustration are speeding

up a once painstaking process — and making it cheaper into the bargain. And advances in animation and three-dimensional visualization are taking scientific illustration beyond the confines of the textbook and the museum wall. In addition to eye-popping graphics for science magazines, textbooks and sundry publications, scientifically accurate illustrations are also needed for legal presentations and forensics. Even more lucratively, the health and pharmaceutical industries have seized on high-end three-dimensional graphics and animation for surgical training, web-based education and marketing.

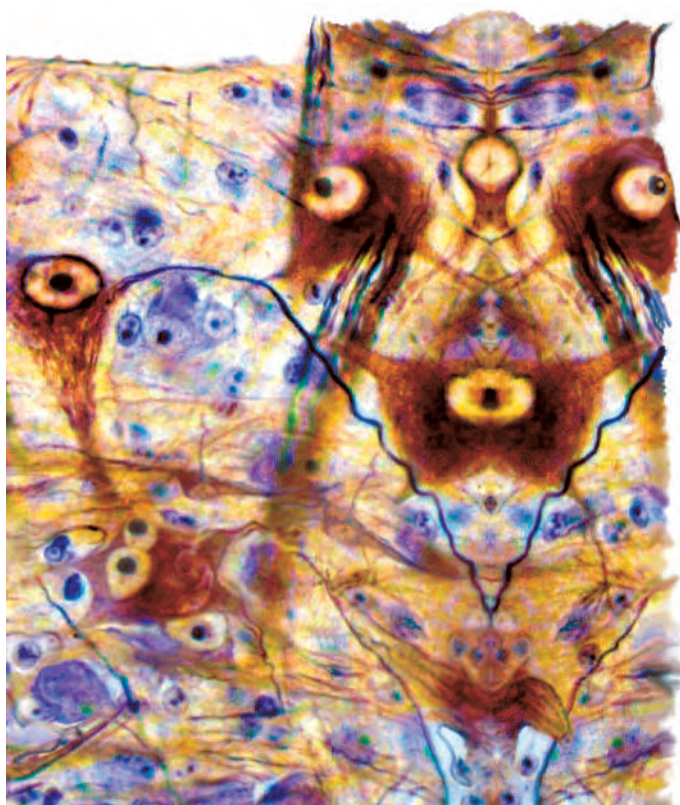
But finding steady work requires more than just artistic flare. Most work is done on short- to mid-term contracts, and on a freelance basis. So beyond artistic skill, fledgling scientific illustrators require Internet savvy and an ability to market themselves. About 40% of the 600-member-strong Association of Medical Illustrators (AMI), based in Cambridge, Massachusetts, are freelance, says the association's president, Marcia Hartsock. Ray Evans, science illustrator and treasurer of the European Association of Medical and Scientific Illustrators (AEIMS), says that networking among colleagues and promoting their work to potential clients are necessary survival skill for illustrators.

Like most editors of science publications, Ed Bell, art director of *Scientific American*, relies on some 20–40 freelancers with different specializations to meet the needs of his multidisciplinary magazine, as it has no staff illustrators. "When illustrations become challenging, these artists can talk to scientists directly," says Bell, noting that a scientific background is important because the artists need to understand what they are trying to depict, as well as how to communicate with writers and editors.

New-wave niches

Although there is still a place for the hand-drawn botanical, anatomical and entomological renderings of yesteryear, most of the current demand in scientific illustration springs from the digital and online worlds. The general market for illustrators is very competitive, says Bell, but there is still room for artists who can handle science to carve themselves a niche — particularly because science is so specialized. Understanding the fundamentals of biology enables a scientific illustrator to translate complex findings into intuitively understood images.

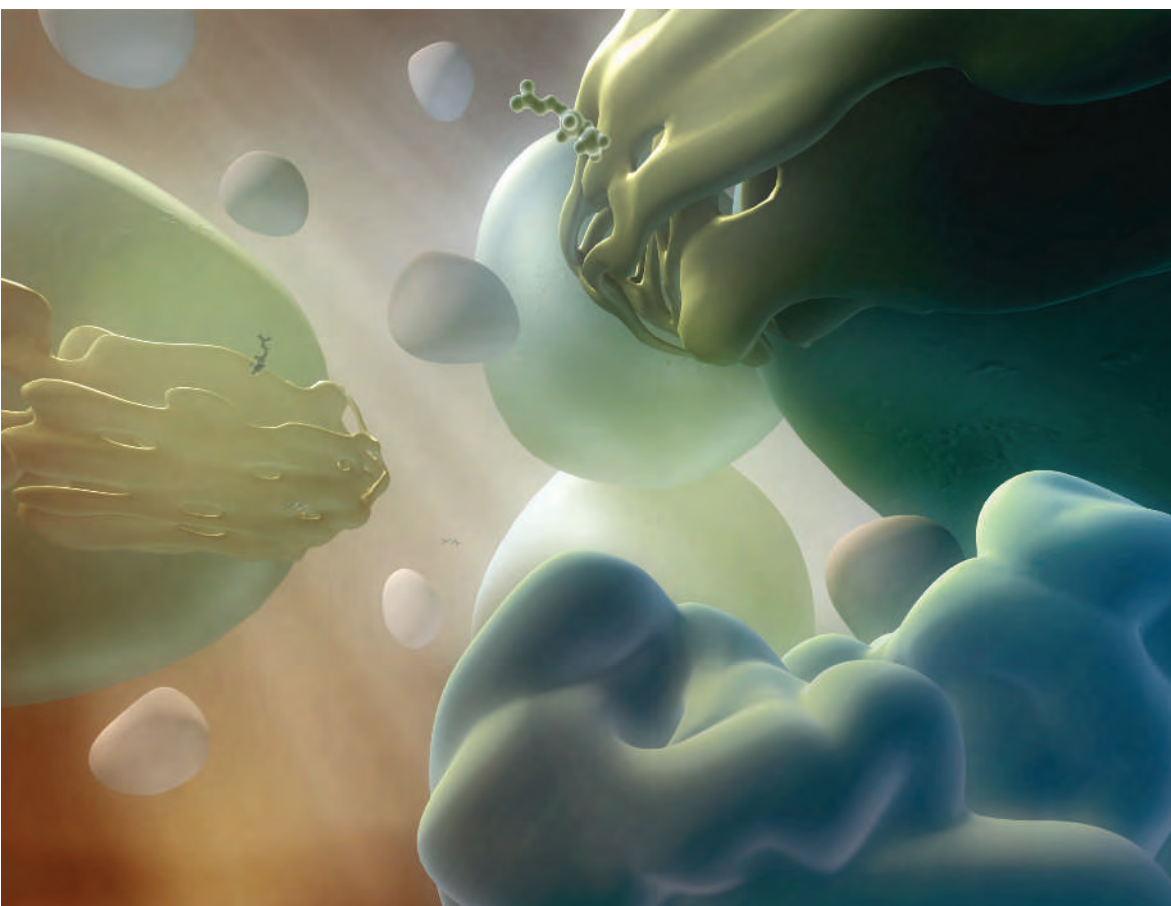
A scientific background can help an artist find a niche in the traditional publishing world, but the biggest growth in work is in professions such as law, marketing, education and entertainment. But to break into these sectors, budding scientific illustrators must expand their palette beyond pen and ink, brush and canvas. Acumen in computer animation and three-dimensional rendering are now prerequisite skills for many jobs, says Bill Glass, director of medical modelling and illustration at Immersion Technologies in Gaithersburg, Maryland. He develops three-dimensional simulators used to train healthcare professionals on the latest medical procedures.



PAINTING THE BRAIN

Formally, Gayle Gross de Núñez describes her art as interpretive histology, but more often than not she refers to it as brainpainting. "The artwork begins with the lab work," she says. In general, Gross de Núñez begins by dissecting brain tissue — in *What the Waters Brought Forth*, above, she has used cells from an embryonic zebrafish brain. Using the tissue as a canvas, and the dyes and reagents from histology as her background paints, she can then add her interpretive flourishes and incorporate her artistic vision.

► www.brainpainter.com



Computer-generated illustration showing cells responsible for breaking down bone.

Once an illustrator combines their artistic skills and scientific background with the latest computational tools, doors to new career corridors may open. Take law, for instance. From medical malpractice to personal injury, illustrations are needed to explain complicated medical concepts to jurors. "Sometimes you really need animation to explain a progression of events," says Bob Shepherd, vice-president of MediVisuals in Richmond, Virginia, adding that sophisticated animations increase jurors' recall of facts.

And forensics, an offshoot of law and criminal justice, is also benefiting by merging sculpture with three-dimensional rendering. This means that a medical illustrator or sculptor can now find work in facial reconstruction, to assist in surgeries, as well as in forensics and archaeological work. "The role of the medical artist is getting much wider," says Caroline Wilkinson, who does facial reconstructions at the Unit of Art in Medicine at the University of Manchester, UK.

An animated option

Perhaps the biggest growth area is animating how drugs work in the body — either for advertising or for educating patients. "Animation is a way to bring terms to life — that's a very powerful educational medium to reinforce learning," says former drug-industry executive Eleanor O'Rangers, now vice-president and group scientific supervisor at Phase Five Communications, a medical-education company in New York. She says that once created, animated pieces can be repackaged and used in several different public-relations capacities.

"The role of the medical artist is getting much wider."

— Caroline Wilkinson

WEB LINKS

Guild of Natural Science Illustrators

♦ www.gnsi.org

Association of Medical Illustrators

♦ medical-illustrators.org

European Association of Medical and Scientific Illustrators (AEIMS)

♦ www.aeims.com

Vesalius Trust

♦ www.vesaliustrust.org

Training opportunities list

♦ fairmanstudios.com/links.htm

Jane Hurd, president of New York-based visual-science agency Hurd Communications, says that increased computing power and the Internet have fuelled a plethora of new venues to showcase high-end three-dimensional animations for the drug industry. "Everything we do now is heavy in cell biology and molecular biology, describing how new pharmaceutical products work," she says. Hurd thinks of her role as content developer as well as animator.

Most scientific illustrators tend to start out proficient in either science or art, and there are increasing opportunities for them to hone the rest of their craft in graduate training programmes, a number of which are now emerging. These tend to take quite different approaches from each other. The University of Manchester's Unit of Art in Medicine, for example, has its roots in developing medical,

surgical or anatomical education materials and teaches art. But students at a two-year programme at the Medical College of Georgia, Augusta, tend to start out with an art and painting background and learn the science.

The science-illustration programme at the University of California, Santa Cruz, is one of the few globally to have a broader purview than medical illustration. It places more emphasis on professional skills. The one-year programme emphasizes learning computer techniques and three-dimensional animation, and is followed by an internship at venues such as the Monterey Bay Aquarium in California or the American Museum of Natural History in New York City.

But how can you afford such training — especially when there are few accredited programmes and virtually no guarantee that completing one will get you a full-time job? Although funded studentships are almost nonexistent in this field globally, the Vesalius Trust, a non-profit organization based in Kildeer, Illinois, funds the training of a few scientist-artists to conduct innovative education and research in the health sciences.

Most industry insiders recommend formal training, although many succeed without it. Gross de Núñez and a colleague took on their first video project without either of them having a clue about medical illustration. They learned three-dimensional animation on the fly, finished the project and kept hustling for work. Whether artist-turned-scientist or scientist-turned-artist, illustrators armed with the right mix of artistic and technological tools can make a masterpiece of their career.

Virginia Gewin is a freelance science writer based in Portland, Oregon.

Omphalosphere: New York 2057

A trip to the Zoo, a visit to the Library.

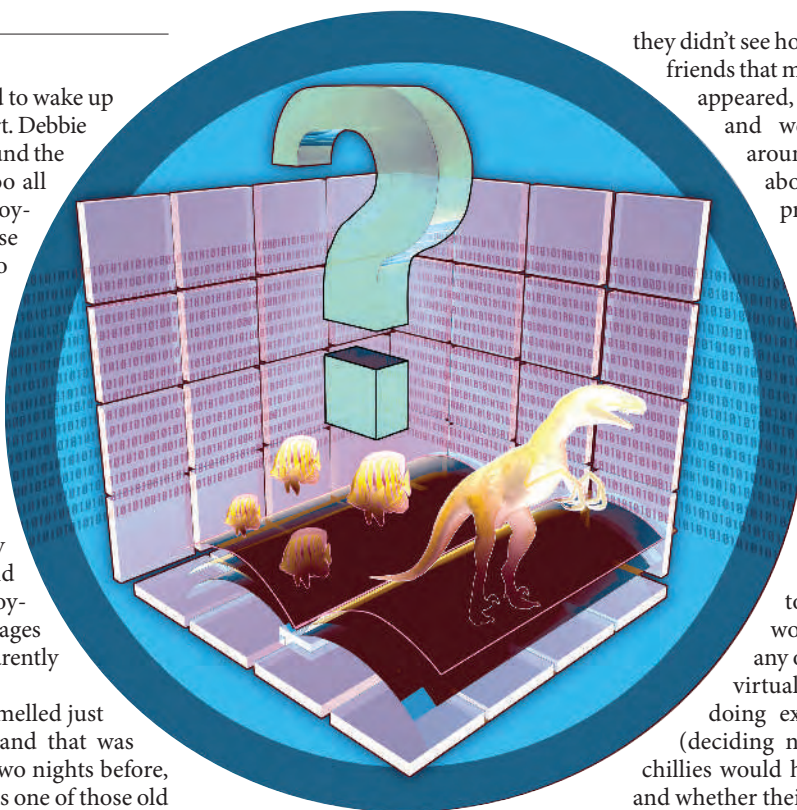
Jack Cohen

I was not at all surprised to wake up and find that my feet hurt. Debbie and I had mooched around the reconstituted Bronx Zoo all of the previous day, enjoying the sights of all those strange animals no longer to be found in this world. We especially enjoyed their sounds, and their particular and peculiar smells. There were clots of people around each of the open cages, enjoying the smells, all so different. So many visitors, of all ages and sexes and cultures, enjoying the old-fashioned cages with their exciting, apparently alive, creatures.

I thought the bears smelled just like Debbie after sex, and that was fresh in my memory. Two nights before, the Library had found us one of those old 'aphrodisiac' books that had recommended the addition of a little pepper — or if you were braver, chillies — to love-making, and we were both regretting this as we walked around the park, despite the considerable volumes of unguents we'd applied to each other. We laughed at the ostriches, and found the elephants nearly as unbelievable as the tyrannosaurs. The aquarium exhibits were simply beautiful, along many axes, from the simply colourful to the ornately grotesque, from the lean lithe predators to the amazingly camouflaged prey, from the coral organisms with their symbiotic algae to their larvae in the busy world of the plankton. Determined to learn some more biology this afternoon.

What was on the agenda today? Ah, Moshe and his research on ancient religions, souls and eternities. A mutual friend, Ali, who taught theology at Columbia University, had found an old, 1930s parable by an early scientist called Haldane, explaining to his audience what eternity was by imagining a Neapolitan street-sweeper child wafted to 'Heaven', and his God offering to have a bird drop a feather on a mountain once a century: "When the mountain is worn away..."

"No," Haldane's child had said (anticipating Pascal's many infinities). "Long before the feathers have had any effect, the



world of birds, feathers and mountains will have been superseded." And so on. "Our successors will think of our time as a pre-Creation eternity: our predecessors found eternity as our time unfolded." A very pretty little parable.

Where, Moshe and Ali wondered, was this 'Heaven', where 'souls' lived, where 'God' was to be found? I observed, as they discussed Plato's ideal plane of existence; and Rucker's Mindscape; and the Judaeo-Christian Creation myth with its Garden of Eden so clearly lifted from the Sumerians; and the question of whether the ancients believed that Adam had a navel, the trees in the Garden had tree-rings, and the rock below had fossils — all the Created were consistent with a long causal but fictional history. The Library found Gosse's *Omphalos* for them, which had propounded that idea to the Victorians as the way of integrating Darwin with Creationist Christianity: God had created an 'old' Universe, and you couldn't catch the Old Guy out.

We all enjoyed coffee and liqueurs as we argued about different planes of existence, what causality was, if there were Laws of Nature that made things go the way they did. "The scientists of the twentieth century believed there were," said Ali, "and

they didn't see how close to their religious friends that made them." Then Debbie appeared, and Ali's current lover, and we all found ourselves around a lunch table talking about the food and its provenance.

I thought how much they — and I — were perpetual students, rehashing all the old arguments and examples with the occasional new parable thrown in by the Library from the literature of the past. I wandered off alone towards the Bronx Zoo again.

I turned on a light rain to match my mood and wondered, again, whether any of the other thousands of virtual New Yorks had me doing exactly the same things (deciding not to experiment with chillies would have been a good idea), and whether their Libraries fed them the same material from the infinite literary and scientific resources before the Uplift.

Then it occurred to me that the Library's literary past might never actually have existed, as unreal as the pre-Creation seasons documented by the tree-rings in the Eden trees. The Library, like the God in *Omphalos*, might have constructed that past so that our civilization seemed rational, with a causal history. The problem for me was that I simply could not imagine a time and place, a three-or-four-dimensional causal universe like the twentieth century so vividly portrayed for us by the Library, where human beings had possessed material 'bodies' that somehow interacted directly with the world around them to implement their wishes. Their supposed belief in 'souls' seemed positively reasonable compared with the idea that human minds had once been tethered to material bodies that needed to eat, to excrete, to heal after material injury. That there had been what they called a 'reality' that people didn't even have the illusion that they controlled. That people didn't always live, as we know that we do, such rich, fulfilled lives in the machinery of the Library.

Jack Cohen is an author and reproductive biologist based in Newent, Gloucestershire, UK.

JACEY